

# SynTab-LLaVA: Enhancing Multimodal Table Understanding with Decoupled Synthesis

Bangbang Zhou<sup>1\*</sup> Zuan Gao<sup>1\*</sup> Zixiao Wang<sup>1</sup> Boqiang Zhang<sup>1</sup>  
Yuxin Wang<sup>1</sup> Zhineng Chen<sup>2</sup> Hongtao Xie<sup>1✉</sup>

<sup>1</sup>University of Science and Technology of China <sup>2</sup>Fudan Univeristy

{bangzhou01, zuangao, wzx99, cyril}@mail.ustc.edu.cn

{htxie, wangyx58}@ustc.edu.cn zhincheng@fudan.edu.cn

## Abstract

Due to the limited scale of multimodal table understanding (MTU) data, model performance is constrained. A straightforward approach is to use multimodal large language models to obtain more samples, but this may cause hallucinations, generate incorrect sample pairs, and cost significantly. To address the above issues, we design a simple yet effective synthesis framework that consists of two independent steps: table image rendering and table question and answer (Q&A) pairs generation. We use table codes (HTML, LaTeX, Markdown) to synthesize images and generate Q&A pairs with large language model (LLM). This approach leverages LLMs high concurrency and low cost to boost annotation efficiency and reduce expenses. By inputting code instead of images, LLMs can directly access the content and structure of the table, reducing hallucinations in table understanding and improving the accuracy of generated Q&A pairs. Finally, we synthesize a large-scale MTU dataset, SynTab, containing 636K images and 1.8M samples costing within \$200 in US dollars. We further introduce a generalist tabular multimodal model, SynTab-LLaVA. This model not only effectively extracts local textual content within the table but also enables global modeling of relationships between cells. SynTab-LLaVA achieves SOTA performance on 21 out of 24 in-domain and out-of-domain benchmarks, demonstrating the effectiveness and generalization of our method. The Code is available at [SynTab-LLaVA](#).

## 1. Introduction

Tables are a crucial form of data representation, commonly used in domains like finance, internet, and academic, etc. Within the AI community, enabling machines to interpret

\*Equal contribution. ✉Corresponding authors.

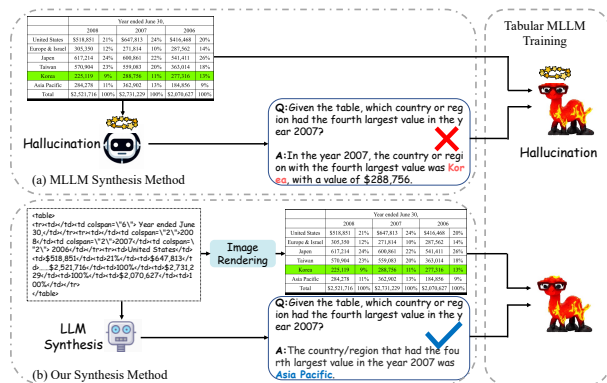


Figure 1. Compared to MLLM synthesis method (a), our synthesis approach (b) alleviates hallucinations in table inputs, enabling the generation of more accurate Q&A pairs and achieving better optimization in subsequent tabular MLLM training. Zoom in for better visualization.

tables has emerged as an active research area [7, 30, 44, 45]. A prominent focus in this field is multimodal table understanding (MTU), which involves answering user questions based on information extracted from table images.

Although recent works [7, 20, 31, 42] improve table understanding, the limited quantity of existing MTU datasets still hampers model performance. Therefore, some other studies [3, 5, 30, 44] focus on annotating more MTU data to enhance model performance. The annotating process begins with obtaining table images, requiring annotators to generate a question and answer (Q&A) pair based on the image. Ultimately, any table understanding sample can be abstracted as a tuple of the form  $\langle \text{image}, \text{question}, \text{answer} \rangle$ . We believe that three key metrics should be considered during the annotation process: efficiency (annotation time per sample), cost (annotation expense per sample), and robustness (sensitivity to variations in input tables). Our investigation reveals that existing methods primarily rely on two approaches: human annotation [3, 5, 16, 30] and multimodal large language

model (MLLM) synthesis [45, 46], neither of which effectively balances the three metrics.

The human annotation process is straightforward. Upon receiving an *image*, annotators generate the corresponding Q&A pairs. Leveraging the strong visual perception abilities, annotators can effectively handle images of varying quality and size, achieving a high degree of robustness. However, the drawback is that samples cannot be generated in parallel. Additionally, hiring human annotators can be quite costly. For instance, creating 2000 sample pairs for TableBench [39] costs \$12,000 in US dollars. The Fig. 1 (a) shows another method [26, 45, 46], which involves using MLLM [28, 34] to synthesize table understanding samples. They design appropriate prompts to guide the MLLM in understanding the table images and generating reasonable Q&A pairs. This approach allows for parallel requests to the MLLM, thus enhancing efficiency. However, the MLLM usually generates a large number of visual tokens from the table images, which increases the synthesis cost. Although TabPedia [45] synthesizes over one million samples for MTU, the dataset has not been made publicly available due to high costs and company copyright restrictions. More critically, the process by which MLLMs handle table images is complex and is easily affected by image quality and structural extraction capabilities, increasing the risk of hallucinations. This includes misinterpreting the structure of complex tables, overlooking certain table content, and incorrectly modeling relationships between cells, *etc.* Such issues lead to inaccuracies in the generated Q&A pairs and reduce the robustness of the synthesis method. As shown in Fig. 1 (a), using these incorrect Q&A pairs to train tabular MLLMs further induces hallucinations, hindering the models from achieving optimal performance.

To address the above issues, this paper proposes a novel method for synthesizing MTU data that balances efficiency, cost, and robustness, shown in Fig. 1 (b). In this method, we introduce decoupling the MTU annotation into two independent steps: table image rendering  $\langle \text{code}, \text{image} \rangle$  and table question and answer pairs generation  $\langle \text{code}, \text{question}, \text{answer} \rangle$ . Here, *code* represents the string sequence (HTML, LaTeX, Markdown) expressed by table images. In table image rendering, we use rendering toolkits to convert codes into images and apply various data augmentation techniques to ensure image diversity. Through this, we obtain 636K  $\langle \text{code}, \text{image} \rangle$  samples. For Q&A pairs generation, we utilize LLM Doubao to treat table *code* as *image* and design a prompt generator to guide the LLM generating Q&A pairs with various question types. Finally, we merge the outputs from the two steps and synthesize a large-scale dataset, **SynTab**, with 1.8M sample pairs  $\langle \text{image}, \text{question}, \text{answer} \rangle$ . Compared to existing datasets, our method has significant superiority on the three key metrics: 1) Lower costs. Thanks to the low price of LLMs and the fact that table codes have fewer

input tokens than images, our method achieves much lower costs during the synthesis process (\$200 vs \$12000 in US dollars). 2) Higher efficiency. Commercial LLMs can handle parallel requests, which significantly enhances the efficiency of data annotation. 3) Better robustness. Our approach directly processes structured table codes, enabling the LLM to achieve higher accuracy and reliability in row-column logic, contextual reasoning, and content consistency. Additionally, this approach reduces visual noise and information loss from table codes. Consequently, compared to MLLM synthesis methods, the Q&A pairs generated by the LLM are more accurate, and coherent, and exhibit significantly fewer hallucinations.

Furthermore, we notice that existing methods [22, 46] utilize  $336 \times 336$  as the input size, while the table images often exceed this size, leading to significant loss of image content information due to excessive resizing.

Therefore, this paper proposes a hybrid multi-resolution multimodal table understanding model, **SynTab-LLaVA**, which encodes local visual information and global structural relationships on high-resolution and low-resolution images, respectively. To validate the effectiveness of the SynTab-LLaVA, we conduct experiments on 24 table understanding benchmarks mentioned in Table-LLaVA [46]. The results demonstrate that our dataset and model effectively enhance table understanding capabilities. Compared to general MLLMs and specialized models for MTU, our approach achieves state-of-the-art performance on 21 benchmarks.

To summarize, our contributions are as follows:

- We propose a novel multimodal table understanding synthesis method that decouples the synthesis steps, ultimately achieving high efficiency, low cost, and high robustness.
- We will open source the first million-level multimodal table understanding synthesis dataset SynTab for tabular multimodal large language model community.
- We propose a multimodal tabular understanding model, SynTab-LLaVA, which effectively enhances model performance in table understanding and achieves SOTA results across multiple benchmarks.

## 2. Related Works

### 2.1. Multimodal Table Understanding Datasets

Existing manually labeled MTU datasets can be categorized into four categories: Table Question Answering (TQA), Table Fact Verification (TFV), Table-to-Text Generation (T2T), and Table Structure Understanding (TSU).

The TQA includes seven datasets related to question answering on table images. The WTQ [30] and FeTaQA [27] respectively require models to generate short and extended answers. HiTab [5] and AIT-QA [17] collect a hierarchical dataset to handle complex tables. For TabMCQ [16], the dataset involves selecting the correct answer from a set of

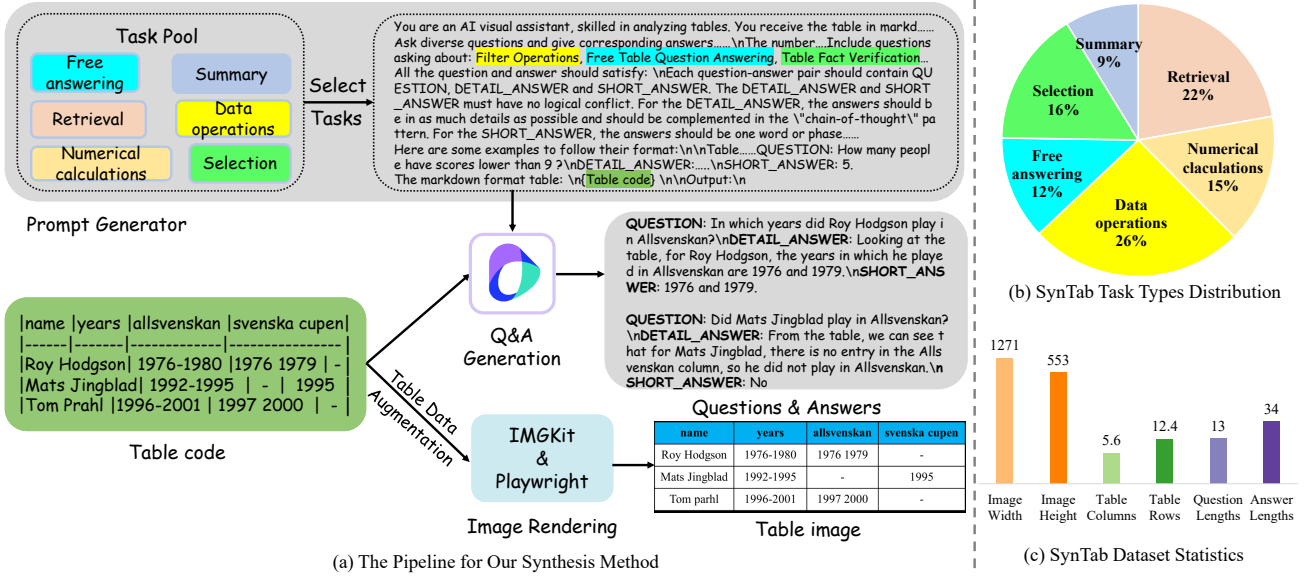


Figure 2. A comprehensive overview of the synthesis framework (a), consisting of two independent steps: table image rendering and table Q&A pairs generation. (b) and (c) present the relevant statistical information of our synthesized dataset.

multiple options provided. Finally, TABMWP [24] and TAT-QA [51] mainly focus on the calculation and reasoning of numerical values in tables.

The TFM category includes TabFact [3], InfoTabs [13], and PubhealthTab [1], which primarily involve retrieving information from tables and comparing it against statements in questions to determine their correctness.

The T2T category includes four datasets. ToTTo [29] and HiTab\_T2T [5] require generating descriptions based on highlighted cell areas in the table. Rotowire [38] provides summaries of sports events, including scoring data for each NBA game. WikiBio [18] is used to generate biographical information about individuals.

The TSU [46] includes Table Size Detection (TSD), Table Cell Extraction (TCE), Table Cell Locating (TCL), Merged Cell Detection (MCD), Row & Column Extraction (RCE), and Table Recognition (TR). It enables the model to enhance the understanding of structural information in table images and OCR capabilities.

Statistics [46] show that these datasets have only 232K training samples, while our SynTab dataset includes 1.8M samples, an order of magnitude larger.

## 2.2. Multimodal Large Language Model

Thanks to the rapid development of LLMs [2, 12, 35, 40], recent studies [19, 22] attempt to integrate multimodal information into LLMs to achieve perception and understanding of visual content, referred to as MLLMs. Despite the good performance in some fields, their input resolutions are relatively low. LLaVA1.5 [22] adopts the resolution of  $336 \times 336$ , which is too small for table images. Excessive resizing of

the images leads to a significant loss of content and detailed information, resulting in incorrect answers from the MLLM. To enhance resolution, some methods [15, 21, 41] attempt to crop a high-resolution image into multiple small non-overlapping patches. While this approach can enhance the input size, it disrupts the spatial relationships and content within the table images, leading to inaccurate understanding. Moreover, this cropping operation generates a large number of visual tokens, which is unacceptable in computation resource-constrained scenarios. In contrast, this paper proposes SynTab-LLaVA that increases the input size to  $1536 \times 1536$  while producing only 576 visual tokens.

## 3. SynTab Dataset

In this section, we provide a detailed description of the synthesis details of SynTab. The synthesis pipeline is illustrated in Fig. 2 (a), which mainly consists of two independent steps: table image rendering and table Q&A pairs generation. In Sec. 3.1 and Sec. 3.2, we explain how to use the sequence representation of table, *code*, including HTML, LaTeX, and Markdown, as intermediaries to synthesize the desired MTU sample pairs  $\langle \text{image}, \text{question}, \text{answer} \rangle$ . In Sec. 3.3, we analyze the advantages of the SynTab dataset.

### 3.1. Table Image Rendering

The purpose of table image rendering is to generate corresponding *image* based on the *code* of the tables. This step helps us obtain the *image* for the MTU samples.

**Code Collection.** We collect 113K and 652K textual tables from the open-source datasets FinTabNet [48] and TableLLama [7], primarily consisting of HTML- and

Markdown-formatted table data. Subsequently, we design detailed scripts to convert the collected data into *codes* that only contain the content and structure of tables. Next, we filter the collected codes, retaining those with a row count between 3 and 60, a column count of at least 3, and fewer than 1400 tokens after tokenization with tiktoken. Ultimately, we filter out 636K tables for subsequent synthesis.

**Image Rendering.** We employ a series of data augmentation techniques aimed at enhancing the visual diversity of table images. These augmentations are applied to various table attributes, such as background color, font size, and font style, to generate a range of visually distinct table images. Additionally, we randomly choose diverse table layouts, including row-only, column-only, or fully gridded separators, to achieve more varied display effects. To further simulate real scenarios, we randomly apply highlighting effects to specific cells within table. This includes continuous blocks of cells, entire rows, or columns, *etc.*, emphasized through shadows or color highlights. By these augmentation techniques, we synthesize visually diverse table images that closely resemble the table images in existing MTU samples.

Through the above process, we synthesize 636K  $\langle \text{code}, \text{image} \rangle$  samples. Following [46], we use these samples in the pre-training of SynTab-LLaVA. Therefore, we refer to the synthesized code-image dataset as **SynTab-Pre**.

### 3.2. Table Question and Answer Pairs Generation

Previous methods [39, 45] request MLLMs or LLMs to automatically generate corresponding Q&A pairs based on the input which limits the diversity of question types. Therefore, after analyzing the types of questions in the MTU samples, we define 6 main question categories and 11 more detailed subcategories to ensure that our synthesized dataset covers a wide range of MTU question types. Additionally, we design a prompt generator to produce suitable prompts that guide the LLM in generating Q&A pairs constrained by specific question types and output formats.

#### 3.2.1 Category Definition

The question types in MTU mainly fall into 6 major categories, encompassing a total of 11 subcategories: retrieval (table retrieval), data operations (counting, ordering, determining range, filtering), numerical calculations (simple numerical calculations, complex calculations), free answering (free table question answering), selection (multiple choice, table fact verification), and summary (table summary).

**Retrieval** involves identifying and extracting specific information or multiple table cells directly from a table. **Data operations** process table data to retrieve values based on the question. These include: counting, ordering, range determination, and filtering. **Numerical calculations** operate quantitative data in tables for analysis. This includes simple numerical addition, subtraction, multiplication, and division

as well as complex mixed operations. **Free answering** requires the answer should integrate both facts and inferences into a coherent sentence in response to the question. **Selection** involves choosing 1 correct option from  $N$  answers, typically  $N = 4$ . Table Fact Verification is a special case where  $N$  is 2, requiring a choice between affirmative or negative responses. **Summary** involves providing a concise and coherent description or caption of the key information presented in a table. We represent detailed descriptions of these 11 subcategories in Supplementary Materials.

#### 3.2.2 Prompt Generator

The prompt generator is designed to generate diverse prompts that guide LLMs to produce Q&A pairs based on question type requests. As shown in the prompt generator in Fig. 2 (a), we first randomly select  $N$  distinct subcategories from the given task pool, where  $N \in [1, 11]$ . The task pool represents the 6 categories defined in last subsection. Next, based on the number of rows and columns in a table, we determine the number of Q&A pairs to be generated for the table. Additionally, we specify the output format in the prompt and provide a structured template for the generated Q&A pairs. Each question corresponds to both a detailed answer and a brief answer to reflect the process of problem-solving. Finally, we assemble the input code with the predefined question types and output format to form a complete prompt, which is then input into the LLM to generate the required Q&A pairs.

Through the above process, we generate a variety of prompt templates, significantly enriching the diversity of question types generated by LLMs. Additionally, detailed answers illustrate the steps taken by the model to solve problems, which is crucial for MLLMs to achieve MTU tasks. We use 636K codes to generate 1.8M sample pairs  $\langle \text{code}, \text{question}, \text{answer} \rangle$ . Subsequently, we combine the images generated in Sec. 3.1 with the Q&A pairs using the *code* as an intermediary to create MTU sample pairs  $\langle \text{image}, \text{question}, \text{answer} \rangle$ , resulting in the dataset **SynTab-SFT** for instruction fine-tuning of SynTab-LLaVA.

### 3.3. SynTab Dataset Analysis

The statistics of SynTab are shown in Fig. 2 (b) and (c). Fig. 2 (b) shows the proportions of the 6 question types in SynTab and (c) presents the statistical information of the samples. It is worth noting that, due to the inclusion of detailed answers, the average length of our answers is 34, which helps the model learn the steps involved in problem-solving.

Compared to previous datasets mentioned in Sec. 2.1, **SynTab** has the following advantages: 1) We are the first to open-source a million-level MTU dataset, containing 636K table images and a total of 1.8M sample pairs, whereas previous MTU datasets only include 82K table images and 232K sample pairs. 2) The table images rendered in Sec. 3.1 are visually highly realistic. We provide visualizations of



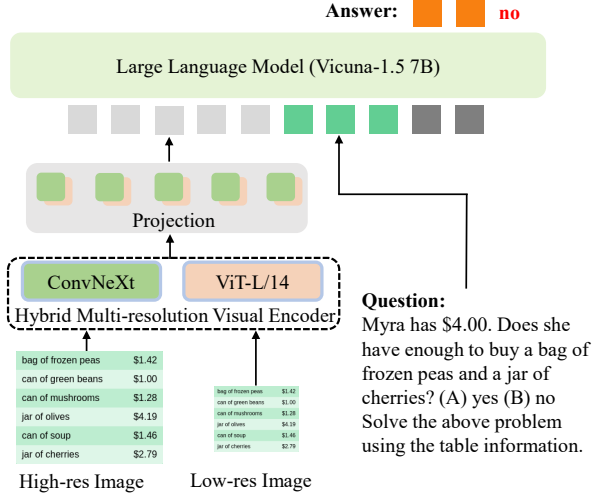


Figure 3. The overall architecture of SynTab-LLaVA.

the synthesized images in the Supplementary Materials. 3) By utilizing the prompt generator, the synthesis of Q&A pairs can cover most question types in MTU, enhancing the generalization capability of MTU models.

## 4. SynTab-LLaVA

As shown in Fig. 3, we present the overview of SynTab-LLaVA. The model framework mainly consists of three parts: the hybrid multi-resolution visual encoder, projection layer, and LLM.

### 4.1. Model Architecture

**Hybrid Multi-resolution Vision Encoder** accepts high-resolution and low-resolution table images as input and uses the corresponding visual encoder to extract relevant local and global information, finally concatenating them in the feature dimension.

**High-Resolution Vision Encoder.** As demonstrated by previous methods [21, 23], high-resolution images allow MLLMs to acquire more visual information, which is also important for table images. Therefore, we adopt a five-stage ConvNeXt [11] to encode high-resolution image inputs. Specifically, given an input image  $I$ , we resize it to a fixed high-resolution size of  $H \times H$ , denoted as  $I_h$ , where  $H$  is set to 1536. Next, the resized image  $I_h$  is fed into ConvNeXt to obtain the feature map  $F_h$ , which is downsampled by a factor of 64, resulting in  $24 \times 24$  visual tokens, each with a dimension of 3072, denoted as  $F_h \in \mathbb{R}^{576 \times 3072}$ .

**Low-Resolution Vision Encoder.** Although ConvNeXt can extract local features from the text regions of table images, it fails to capture the spatial contextual relationships between cells. Therefore, we use ViT-L/14 [33] to model the global spatial relationships in table images. Specifically, we resize the image  $I$  to  $336 \times 336$ , and after processing with ViT-L/14, we obtain 576 visual tokens, each with a

dimension of 1024, denoted as  $F_l \in \mathbb{R}^{576 \times 1024}$ .

Afterward, we merge the  $F_h$  and  $F_l$  along the feature dimension to form a hybrid resolution visual feature:  $F = \text{concat}([F_h, F_l], \text{dim} = -1)$ . Each token in  $F$  contains not only local visual information from the corresponding image region but also the global spatial and structure relationships of the table cells. Furthermore, this fusion approach ensures that the feature  $F$  still retains 576 tokens, thus avoiding additional resource consumption in the self-attention of the LLM. We employ a two-layer MLP as projection layer to convert the visual features into the embedding space of the LLM. Finally, we combine the visual features with the embedded textual features and input them into Vicuna [6] to generate answers.

### 4.2. SynTab-LLaVA Training

Similar to the training strategies of other MLLMs [22, 45, 46], we adopt a two-stage process for training SynTab-LLaVA, consisting of pre-training followed by instruction fine-tuning.

**SynTab-LLaVA Pre-training.** The MTU needs to understand the structural information and textual content in the table images and reason correct answers to the questions posed. So we select table recognition as the main goal for the pre-training, which requires to output of the corresponding *code* based on the input table *image*. In summary, our pre-training data consists of table recognition samples (SynTab-Pre 636K, MMTab-Pre [46] 150K) and general image-text alignment samples (LLaVA1.5-Pre [22] 558K).

**Multimodal Table Understanding Fine-tuning.** Previous MLLMs [21, 22, 45] employ a variety of datasets, including captions, visual question answering, and optical character recognition to achieve general multimodal understanding. However, these datasets are not particularly effective for the MTU task. Therefore, we only use MTU-related datasets, which include the SynTab-SFT with 1.8M samples and the total 232K training samples mentioned in Sec. 2.1.

## 5. Experiments

### 5.1. Implementation Details

**Model Configuration.** We conduct experiments based on the well-trained ViT-L/14 [33], ConvNeXt [11], and Vicuna-1.5 7B [6]. We employ a cosine schedule with a one-cycle learning rate strategy and utilize a warm-up for the first 3% of the training process. During the pre-training, we set the learning rate for the projection layer to  $1e-3$ , and the batch size to 256. In the fine-tuning, we apply LoRA to Vicuna with a rank of 128 and learning rate of  $2e-4$ . As for projection layer, the learning rate is set to  $2e-5$ .

**Evaluation Benchmarks and Metrics.** To evaluate the effectiveness of SynTab on MTU, we apply multiple benchmarks to each MTU subtask. The TQA task consists of

Method	LLM	Res.	Table Question Answering							Table Fact Verification			Table-to-Text Generation			
			TABMWP	WTQ	HiTab	TAT-QA	FeTaQA	AIT-QA	TabMCQ	TabFact	InfoTabs	PubHealthTab	ToTTo	HiTab_T2T	Rotowire	WikiBIO
			Acc.	Acc.	Acc.	Acc.	BLEU	Acc.	Acc.	Acc.	Acc.	Acc.	BLEU	BLEU	BLEU	BLEU
Open-source MLLM																
LLaVa v1.5 <sup>†</sup> [22]	Vicuna-1.5 7B	336	6.05	1.24	2.03	2.97	8.24	-	-	18.9	28.31	-	6.40	2.07	1.92	2.34
Vary-toy <sup>†</sup> [37]	Qwen 1.8B	1024	4.42	7.96	3.42	8.81	2.44	9.39	-	6.33	6.98	-	0.70	0.27	0.46	0.37
Monkey <sup>†</sup> [21]	Qwen 7B	896	13.26	19.07	6.41	12.31	3.41	-	17.89	22.56	22.11	18.89	3.50	1.12	0.03	2.77
Docow11.5 [14]	LLaMA2 7B	448	11.41	26.80	11.10	12.44	3.58	46.18	3.21	27.67	28.74	28.42	7.60	3.78	0	0
IX2.5 [43]	InternLM2 7B	560	26.65	38.09	15.29	19.43	10.29	51.08	18.51	13.72	8.63	9.17	5.50	2.02	2.69	3.04
Ovis1.5 [25]	LLaMA-3 8B	384	27.37	15.00	6.41	15.67	11.38	13.31	12.44	8.28	25.04	18.18	5.20	2.41	2.34	3.23
InternVL2 8B [4]	InternLM2.5 7B	448	16.51	25.09	7.68	13.99	12.14	23.28	40.04	28.32	36.63	37.23	6.10	1.53	1.57	2.26
Qwen2-VL 7B [36]	Qwen2 7B	-	26.04	37.38	25.19	22.67	10.77	67.12	50.73	16.29	39.67	32.6	15.10	2.96	2.50	4.02
TabPedia [45]	Vicuna-1.5 7B	1920 × 2560	12.27	20.37	1.22	9.71	12.51	17.22	0.87	28.84	9.20	21.01	2.60	1.22	0.03	1.11
Table-LLaVA 7B [46]	Vicuna-1.5 7B	336	57.78	18.43	10.09	12.82	25.60	5.48	44.51	59.85	65.26	51.03	23.00	9.74	10.46	9.68
Table-LLaVA 13B [46]	Vicuna-1.5 13B	336	59.77	20.41	10.85	15.67	28.03	6.06	51.51	65.00	66.91	48.46	24.10	10.40	8.83	9.67
Closed-source MLLM																
GPT-4V Low-res <sup>†</sup>	Unknown	512	60.00	22.50	9.50	19.50	9.26	19.00	66.00	45.50	58.50	59.50	-	1.85	3.89	1.55
GPT-4V High-res <sup>†</sup>	Unknown	768 × 2000	60.50	48.00	27.50	32.50	11.04	62.50	66.00	45.50	65.60	67.00	-	2.98	4.23	1.94
Ours																
SynTab-LLaVA	Vicuna-1.5 7B	1536	88.30	39.59	35.66	51.94	35.45	28.57	70.55	70.78	69.42	68.02	34.60	14.16	14.11	14.06

Table 1. Evaluation results on TQA, TFV, and T2T. Blue highlights indicate the best result achieved for each benchmark. <sup>†</sup> means the results are cited from Table-LLaVA. For other models, we follow Table-LLaVAs evaluation scripts to fairly test.

seven evaluation sets: WTQ, FeTaQA, HiTab, AIT-QA, TabMCQ, TAT-QA, and TABMWP. For the TFV task, TabFact, InfoTab, and PubHealthTab are tested. As for the T2T task, it includes four benchmarks: ToTTo, HiTab.T2T [5], Rotowire, and WikiBIO. The final task, TSU, includes six benchmarks: TSD, TCE, TCL, MCD, RCE, and TR. Among these benchmarks, some samples are out-of-domain whose data do not appear in training. We refer to these test samples as TSD\_OOD, TCE\_OOD, TCL\_OOD, and RCE\_OOD, respectively. For FeTaQA in TQA and all benchmarks in T2T, we use BLEU as the evaluation metric. For other benchmarks in TQA and TFV, accuracy is applied to assess the model’s performance. For TSU tasks, we follow the methods [46, 49] to calculate the metrics. For TSD, we calculate the accuracy based on the predicted rows and columns. For TCE and TCL, we compute accuracy at the cell level. For MCD and RCE, we utilize the F1 score as the evaluation metric. For TR, we use Tree-Edit-Distance-based Similarity to score the recognized strings.

## 5.2. Quantitative Results

We report the quantitative results of the previous MLLM methods on the tasks TQA, TFV, T2T, and TSU, and compare them with our proposed SynTab-LLaVA.

**Performance on TQA.** We compare SynTab-LLaVA with previous MLLM methods, including the open-source models Qwen2-VL [36] and Table-LLaVA [46], as well as the closed-source GPT-4V. As shown in Tab. 1, SynTab-LLaVA achieves top performance, ranking first or second across 6 benchmarks compared to other MLLMs. Notably, there are significant performance gains on the TABMWP and TAT-QA, which are specifically designed for table numerical calculation. Our method shows an average improvement of 23.6% and 32.4% over GPT-4V and Table-LLaVA, respec-

tively. This enhancement can be attributed to the diverse numerical calculation problems included in SynTab. On WTQ, our model also outperforms other MLLMs, ranking just below GPT-4V. This may be due to GPT-4V being trained on the WTQ test sets and having a larger model size. For AIT-QA, our model achieves an accuracy of 28.57, significantly lower than Qwen2-VL and IX2.5. We attribute this to our method of padding the image to a square based on its longest edge before resizing. With an average width of 3159 and height of 600, the extensive padding in the height reduces the model’s ability to interpret the table content. In contrast, Qwen2-VL and IX2.5 use image slicing, avoiding excessive padding. To validate this hypothesis, we compare two MTU models, TabPedia and Table-LLaVA, which also adopt the same strategy as ours, and find that their performance on AIT-QA is similarly low, at 17.22 and 6.06, respectively. Compared to these two models, our approach shows significant improvement in AIT-QA, indicating that the synthesized data can effectively enhance the performance of TQA tasks.

**Performance on TFV.** Table Fact Verification tests the model’s comprehensive abilities in information extraction, reasoning, multimodal alignment, and knowledge inference from table images. As shown in Tab. 1, earlier models like LLaVA 1.5 and Monkey [21, 22] can recognize text content within images but lack understanding of table structure, which results in a significant performance gap compared to SynTab-LLaVA. This observation also highlights the distinction between table images and natural scene images, as the unique structural information of tables can critically impact the model’s judgment of textual statements. These results shown in Tab. 1 demonstrate the effectiveness of our SynTab. Training it jointly with the manually annotated MTU dataset enhances the MTU model’s performance on TFV task.

**Performance on T2T.** As shown in Tab. 1, existing meth-

Method	LLM	Res.	TSD		TSD.OOD		TCE	TCE.OOD	TCL	TCL.OOD	MCD	RCE		RCE.OOD		TR		
			Row Acc.	Col. Acc.	Row Acc.	Col. Acc.	Acc.	Acc.	Acc.	Acc.	F1	Row F1	Col. F1	Row F1	Col. F1	HTML TEDS	Markdown TEDS	LaTeX TEDS
Open-source MLLM																		
LLaVA v1.5 [22]	Vicuna-1.5 7B	336	0.80	2.50	2.40	-	0.22	-	0.62	0.93	1.26	1.66	4.13	-	-	12.88	10.74	1.55
Vary-toy [37]	Qwen 1.8B	1024	1.30	2.20	-	-	1.96	-	0.73	-	0.52	2.01	2.38	-	-	10.13	12.72	11.67
Monkey [21]	Qwen 7B	896	0.80	0.60	-	-	1.46	0.76	1.31	-	0.67	3.89	4.53	4.29	-	21.96	13.29	4.54
Docowl1.5 [14]	LLaMA2 7B	448	0.20	1.50	0.40	0.80	1.00	1.80	0	0	0	2.40	4.60	0.60	1.30	7.40	5.30	0
IX2.5 [43]	InternLM2 7B	560	1.60	8.00	2.40	12.40	3.66	3.80	1.25	3.13	2.30	0.17	0.09	0	0	26.91	62.84	31.45
Ovis1.5 [25]	LLaMA-3 8B	384	5.40	13.00	7.60	16.40	4.69	5.64	2.39	3.33	2.42	13.19	20.30	26.43	37.84	44.31	67.21	43.02
InternVL2 8B [4]	InternLM2.5 7B	448	7.70	34.10	12.40	40.80	11.59	12.36	4.86	9.45	0.79	1.46	9.86	1.11	6.95	48.15	70.00	49.21
Qwen2-VL 7B [36]	Qwen2 7B	-	4.70	15.30	5.20	20.40	9.10	8.35	3.86	6.13	1.03	17.14	23.07	21.52	30.39	32.95	75.97	44.91
TabPedia [45]	Vicuna-1.5 7B	1920 × 2560	2.80	10.20	4.80	12.40	0.16	0.11	0	0	0	1.53	3.73	1.62	1.21	0	12.91	0
Table-LLaVA 7B [46]	Vicuna-1.5 7B	336	33.10	33.20	25.20	16.40	19.45	11.28	29.31	26.10	17.14	31.43	37.93	21.97	18.14	50.24	44.82	46.11
Table-LLaVA 13B [46]	Vicuna-1.5 13B	336	34.40	27.60	31.60	14.80	19.53	11.38	29.68	26.17	16.52	31.07	41.49	21.94	18.67	51.44	46.00	46.50
Closed-source MLLMs																		
GPT-4V Low-res	Unknown	512	6.00	24.00	8.00	15.00	3.57	10.29	14.41	17.73	2.12	30.32	56.86	27.69	50.36	41.55	45.74	34.46
GPT-4V High-res	Unknown	768 × 2000	12.50	46.00	19.00	38.00	9.75	14.36	23.38	27.91	3.50	26.44	43.17	48.52	57.14	48.58	60.58	37.66
Ours																		
SynTab-LLaVA	Vicuna-1.5 7B	1536	56.20	82.40	51.60	62.80	50.29	44.90	60.80	51.33	52.93	54.16	73.33	51.73	55.55	74.58	79.17	76.42

Table 2. Evaluation results on TSU. For any benchmark, we consider SynTab-LLaVA to be the best only when it outperforms other methods across all metrics.

SynTab-Pre	SynTab-SFT	TQA	TFV	T2T	TSU
-	-	41.03	65.62	17.60	54.02
✓	-	41.52	65.69	18.10	55.71
	△	+0.49	+0.07	+0.50	+1.69
-	✓	45.00	68.07	18.88	58.15
	△	+3.97	+2.45	+1.28	+4.13
✓	✓	45.89	68.93	19.07	58.56
	△	+4.86	+3.31	+1.47	+4.54

Table 3. The Effectiveness of SynTab. △ represents the performance gap compared to the baseline.

ods generally perform poorly on these benchmarks. This is because the available training data for these tasks is limited, and most general MLLMs have not been trained on these datasets. Table-LLaVA, however, shows improved performance as it utilizes the T2T training set. In contrast, we add synthetic T2T data for training. Although this data does not fully align with the question types of these three benchmarks, it still aids the T2T task and achieves SOTA performance across all test sets.

**Performance on TSU.** This task primarily assesses the MLLM’s ability to understand basic structural information of table images and its OCR perception capabilities. Leveraging the 636K synthesized table images generated in Sec. 3.1 for pre-training and hybrid multi-resolution vision encoder significantly enhances the models comprehension of table structures and content, leading to substantial performance improvements. As shown in Tab. 2, SynTab-LLaVA achieves SOTA performance on all in-domain benchmarks, demonstrating the effectiveness of SynTab-Pre. Furthermore, while the model’s performance drops on out-of-domain benchmarks, it still significantly outperforms other methods. This suggests that the diversity of rendered images in our SynTab has greatly improved the model’s generalization ability on

ConvNeXt	ViT-L/14	TQA	TFV	T2T	TSU
-	✓	35.78	64.70	14.18	44.46
✓	-	45.12	68.18	18.76	57.91
	△	+9.34	+3.48	+4.58	+13.45
✓	✓	45.89	68.93	19.07	58.56
	△	+10.11	+4.23	+4.89	+14.10

Table 4. The Impact of Hybrid Multi-resolution Vision Encoder.

the TSU task.

### 5.3. Ablation Studies

In this section, we explore the effectiveness of the proposed SynTab and SynTab-LLaVA separately. To reduce training time, we randomly select 20% of the SynTab-Pre and SynTab-SFT data for ablation study.

**Effectiveness of SynTab.** We conduct the following experiments to verify that the synthesized SynTab can assist the manually annotated data in enhancing the performance of MTU models and the results are shown in Tab. 3. The first two columns indicate whether SynTab is used in the pre-training and fine-tuning stages, respectively. The second row indicates that training is conducted using only manually annotated data, resulting in relatively low average performance across the four tasks. When SynTab-Pre is introduced during pre-training, a significant improvement is observed in the TSU compared to other tasks. This enhancement stems from the table recognition task during pre-training, which strengthens the models basic understanding of table structures, hence making the additional training data particularly beneficial for the TSU task. In the seventh row, both SynTab-Pre and SynTab-SFT are used, leading to a marked overall performance improvement. These results indicate the effectiveness of our synthesized data. Despite occasional labeling errors in SynTab, its combination with human-annotated data proves beneficial in optimizing the MTU model, demonstrating a

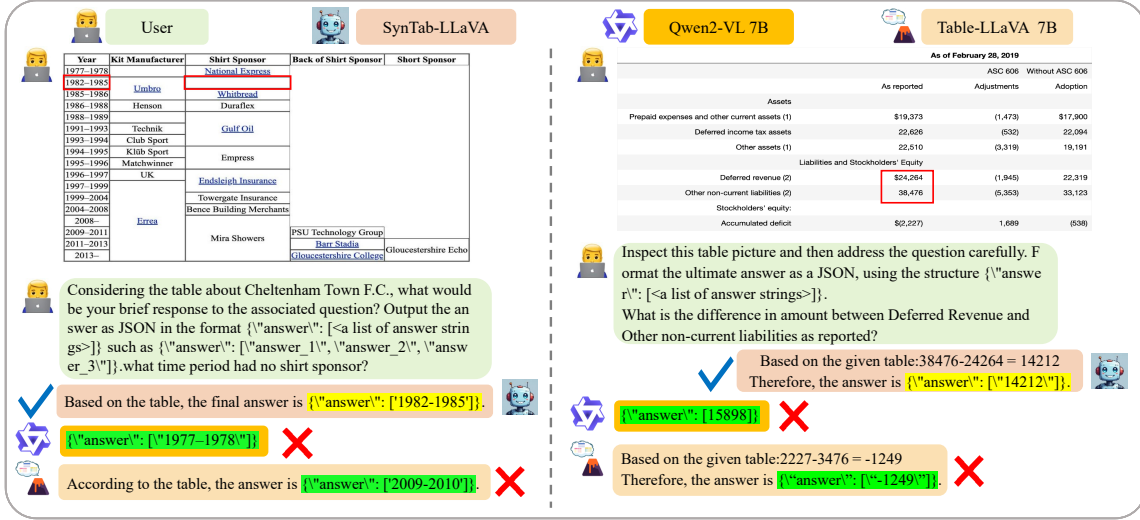


Figure 4. Visualization on WTQ and TAT-QA samples, where yellow and green highlights represent the correct and incorrect answers of the model, respectively. The red rectangular boxes indicate the table cells relevant to the question.

positive effect.

**Impact of Hybrid Multi-resolution Vision Encoder.** As shown in Tab. 4, we experiment with different vision encoders to evaluate their performance. In SynTab-LLaVA, we propose a hybrid multi-resolution vision encoder to capture both global structural information and local cell OCR content at different resolutions. The high-resolution encoder, ConvNeXt, effectively captures rich textual information within each cell region, enhancing OCR perception [8–10, 32, 47, 50]. In contrast, the low-resolution encoder ViT-L/14 leverages global self-attention to learn structural features. The results in this table demonstrate that the synergy between these two vision encoders significantly enhances the model’s ability to understand both the structure and content of table images.

#### 5.4. Visualization

In this part, we visualize the outputs of general MLLM Qwen2-VL and tabular MLLM Table-LLaVA on two samples from WTQ and TAT-QA, as shown in Fig. 4. In the left figure, due to the several merged cell areas in table, both Qwen2-VL and Table-LLaVA fail to accurately associate the year with the corresponding shirt sponsors, resulting in incorrect answers. Table-LLaVA’s response is particularly inaccurate, even producing a year not present in the table, highlighting its limitations in textual extraction of table content. In contrast, our method accurately extracts text from each cell in complex table and effectively establishes associations between cells, demonstrating its effectiveness. The above visualizations effectively validate that SynTab-LLaVA can accurately understand table images and provide precise answers based on user questions. Additional visualizations

are provided in the Supplementary Material.

#### 6. Limitation

Although SynTab-LLaVA significantly enhances MTU performance, several limitations remain to be discussed. First, the synthetic dataset SynTab contains some erroneous annotations, which may hinder the model from learning optimal parameter states. Second, both manually labeled datasets and SynTab are used for academic research, primarily covering data from sources like Wikipedia, web screenshots, and financial reports, *etc.* However, in industrial scenarios, table images are often subject to various conditions such as rotation, occlusions, deformations, and perspective transformations, leading to substantial performance degradation in MTU models.

#### 7. Conclusion

This paper analyzes the shortcomings of existing MTU data annotation methods, mainly including hallucinations and high costs, and presents a novel synthesis approach that is low-cost, efficient, and robust. Specifically, we decouple the synthesis of MTU samples into two independent steps: table image rendering and table question and answer pairs generation. Using this approach, we generate the large-scale, MTU dataset SynTab, which includes 636K images and 1.8M training samples, with the total cost under \$200. Furthermore, we propose SynTab-LLaVA, a hybrid multi-resolution multi-modal table understanding model designed to improve the comprehension of both textual and structural information within table images. Experimental results demonstrate that our approach significantly outperforms existing MLLMs and the powerful GPT-4V across multiple benchmarks.



**Acknowledgments:** This work is supported by the National Nature Science Foundation of China (62425114, 62121002, U23B2028, 62232006). This research is support by the Supercomputing Center of the USTC. We acknowledge the support of GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

## References

- [1] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1–16, Seattle, United States, 2022. Association for Computational Linguistics. 3
- [2] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, and Xiaoyi Dong. Internlm2 technical report, 2024. 3
- [3] Wenhui Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyu Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*. 1, 3
- [4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6, 7
- [5] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1110, 2022. 1, 2, 3, 6
- [6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6, 2023. 5
- [7] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: table understanding through representation learning. *Proceedings of the VLDB Endowment*, 14(3):307–319, 2020. 1, 3
- [8] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. SVTR: scene text recognition with a single visual model. In *IJCAI*, pages 884–890, 2022. 8
- [9] Yongkun Du, Zhineng Chen, Yuchen Su, Caiyan Jia, and Yu-Gang Jiang. Instruction-guided scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [10] Zuan Gao, Yuxin Wang, Yadong Qu, Boqiang Zhang, Zixiao Wang, Jianjun Xu, and Hongtao Xie. Self-supervised pre-training with symmetric superimposition modeling for scene text recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024. 8
- [11] Chunjiang Ge, Sijie Cheng, Ziming Wang, Jiale Yuan, Yuan Gao, Jun Song, Shiji Song, Gao Huang, and Bo Zheng. Con-vllava: Hierarchical backbones as visual encoder for large multimodal models. *arXiv preprint arXiv:2405.15738*, 2024. 5
- [12] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, and Hao Yu. Chatglm: A family of large language models from glm-130b to glm-4 all tools, 2024. 3
- [13] Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Sriumar. INFOTABS: Inference on tables as semi-structured data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2309–2324, Online, 2020. Association for Computational Linguistics. 3
- [14] Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3096–3120, 2024. 6, 7
- [15] Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024. 3
- [16] Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. Tabmcq: A dataset of general knowledge tables and multiple-choice questions. *arXiv preprint arXiv:1602.03960*, 2016. 1, 2
- [17] Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, et al. Ait-qa: Question answering dataset over complex tables in the airline industry. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2022. 2
- [18] Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213, 2016. 3
- [19] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3
- [20] Yiren Li, Zheng Huang, Junchi Yan, Yi Zhou, Fan Ye, and Xianhui Liu. Gfte: graph-based financial table extraction. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*, pages 644–658. Springer, 2021. 1
- [21] Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26763–26773, 2024. 3, 5, 6, 7

- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3, 5, 6, 7
- [23] Yuliang Liu, Biao Yang, Qiang Liu, Zhang Li, Zhiyin Ma, Shuo Zhang, and Xiang Bai. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024. 5
- [24] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*, 2023. 3
- [25] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv:2405.20797*, 2024. 6, 7
- [26] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. ChartAssistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7775–7803, Bangkok, Thailand and virtual meeting, 2024. Association for Computational Linguistics. 2
- [27] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022. 2
- [28] OpenAI. Chatgpt: An ai language model, 2024. Accessed: 2024-10-31. 2
- [29] Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, 2020. 3
- [30] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, 2015. 1, 2
- [31] Devashish Prasad, Ayan Gadpal, Kshitij Kapadni, Manish Visave, and Kavita Sultanpure. Cascadetabnet: An approach for end to end table detection and structure recognition from image-based documents. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 572–573, 2020. 1
- [32] Yadong Qu, Yuxin Wang, Bangbang Zhou, Zixiao Wang, Hongtao Xie, and Yongdong Zhang. Boosting semi-supervised scene text recognition via viewing and summarizing. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. 8
- [33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [34] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2
- [35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3
- [36] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 6, 7
- [37] Haoran Wei, Lingyu Kong, Jinyue Chen, Liang Zhao, Zheng Ge, Jinrong Yang, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pages 408–424. Springer, 2025. 6, 7
- [38] Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017. 3
- [39] Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xinrun Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. *arXiv preprint arXiv:2408.09174*, 2024. 2, 4
- [40] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, and Fei Huang. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 3
- [41] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, 2023. 3
- [42] Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. Tabert: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. 1
- [43] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 6, 7
- [44] Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. Tablellama: Towards open large generalist models for tables. In *Proceedings of the 2024 Conference of the North American*

*Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6024–6044, 2024. 1

- [45] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *Advances in Neural Information Processing Systems*, 37:7185–7212, 2024. 1, 2, 4, 5, 6, 7
- [46] Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2, 3, 4, 5, 6, 7
- [47] Tianlun Zheng, Zhineng Chen, Shancheng Fang, Hongtao Xie, and Yu-Gang Jiang. Cdistnet: Perceiving multi-domain character distance for robust text recognition. *International Journal of Computer Vision*, 132(2):300–318, 2024. 8
- [48] Xinyi Zheng, Doug Burdick, Lucian Popa, Peter Zhong, and Nancy Xin Ru Wang. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. *Winter Conference for Applications in Computer Vision (WACV)*, 2021. 3
- [49] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020. 6
- [50] Bangbang Zhou, Yadong Qu, Zixiao Wang, Zicheng Li, Boqiang Zhang, and Hongtao Xie. Focus on the whole character: discriminative character modeling for scene text recognition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1762–1770, 2024. 8
- [51] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. 3