NTCIR-18: Proceedings of the 18th NTCIR Conference on Evaluation ofInformation Access Technologies, June 10-13, 2025, Tokyo, Japan
DOI: https://doi.org/10.20736/0002002099

STMK24 NTCIR18 U4 Table QA Submission

Hayato Aida
Stockmark
Japan
hayato.aida@stockmark.co.jp

Kosuke Takahashi
Stockmark
Japan
kosuke.takahashi@stockmark.co.jp

Takahiro Omi
Stockmark
Japan
takahiro.omi@stockmark.co.jp

Abstract

This paper describes the methods, results, and analysis of team STMK24 for the NTCIR-18 U4 Table Question Answering (TQA) task. STMK24 approaches TQA as a Visual Document Understanding task, and convert every table into three complementary modalities—image, text, and layout. To simply comprehend the structures of the tables, our model is trained to infer the cell IDs of the tables, and the cell values are automatically extracted through rule-based conversion. We investigated the impact of each modality on Table QA performance and confirmed that the model achieves high cell ID inference accuracy when utilizing all modalities.

Keywords

Table QA, VQA, LVLM, Multimodal

Team Name

STMK24

Subtask

Table Question Answering (Japanese)

1 Introduction

A table is a structured representation of data, and in the business settings, it is provided in various formats. In the NTCIR-18 U4 task [5], tables extracted from Japanese securities reports are provided in HTML format by the Japanese Financial Services Agency. In this paper, we present our proposed system for the Table Question Answering (TQA) task of U4, where questions regarding specific tables and cells are posed, and answers must be generated based on information contained within those tables.

Tables in business contexts represented in diverse formats such as HTML, markdown, CSV, images, and even PDF files, which complicates fully automatic processing. With the recent advancements of Large Vision-Language Models (LVLMs), images and text can be jointly processed, enabling effective and format-agnostic table interpretation through the integration of image modality.

In this study, we propose a multimodal approach by modifying the LVLM architecture to accommodate not only image inputs but also text and layout modalities. Specifically, we convert HTML tables into distinct image, text, and layout representations, enabling our modified LVLM to effectively integrate these modalities for enhanced table comprehension.

Our experimental results demonstrate the effectiveness of this multimodal approach, achieving an accuracy of 95.36% in the cell ID inference task. Furthermore, our modality contribution analysis reveals the critical roles of layout and textual information, highlighting their importance in accurately understanding and interpreting tables. This analysis provides valuable insights into the optimal use of LVLMs for automated table understanding in diverse business settings.

2 Related Work

Visual document understanding encompasses the extraction and interpretation of information from document images to answer relevant queries. Within the area of visual document understanding, Table QA specifically focuses on comprehending tabular information contained within documents. Various benchmarks have been established for assessing performance of the down-streaming tasks in this field. For instance, DocVQA [10] involves extracting and understanding textual and visual content from diverse document images to respond accurately to posed questions. Similarly, datasets like CORD [11] and FUNSD [4] focus on specialized tasks such as receipt understanding and form information extraction from scanned documents, respectively. These datasets leverage multimodal information, including visual features, textual content, and spatial layouts, which are crucial elements for accurate table comprehension.

Recent advancements in transformer-based architectures have further improved multimodal document understanding. LayoutLMv3 [3], for example, incorporates visual, textual, and spatial modalities to achieve state-of-the-art results across various document understanding tasks.

More recently, LVLMs have demonstrated their potential for document understanding tasks by leveraging the strong text-processing capabilities of high-performing LLMs. Models such as LLaVA [7] and LLaVA-OV [6] have shown promising capabilities in general multimodal tasks. Specifically, Qwen2-VL [13] has achieved state-of-the-art performance on the DocVQA benchmark.

LayoutLM[14], on the other hand, is specifically tailored for document image understanding. It extends traditional encoder-based language models by incorporating not only textual content but also the spatial layout of documents. Based on the Transformer architecture, LayoutLM integrates token embeddings with 2D positional embeddings that represent the coordinates of text within a document. This approach allows the model to capture the structural information essential for understanding complex documents, such as forms or receipts. Subsequent versions, like LayoutLMv2[15] and LayoutLMv3[3], further enhance this capability by incorporating actual image embeddings alongside text and layout information, enabling a more comprehensive understanding of documents that include both textual and visual elements.

Additionally, models that leverage only textual and layout information have also been explored. LayTextLLM[8] focuses on integrating textual content with spatial layout information. It achieves this integration by mapping each bounding box to a single embedding and interleaving it with the corresponding text. This approach

efficiently addresses sequence length issues and leverages the autoregressive traits of large language models (LLMs), enabling effective document understanding without relying on explicit visual inputs.

With the advancement of large language models (LLMs) utilizing Transformer decoders with layout awareness, QA datasets incorporating layout information have also been proposed. LayoutLLM[1, 9] introduces a QA dataset that leverages document images along with the text and layout information within them, while also presenting a baseline architecture. The construction of this dataset has the potential to enhance LLMs' ability to comprehensively understand text, layout, and images.

Motivated by these developments, we build upon existing LVLM frameworks by: (1) adopting an LVLM capable of efficiently processing high-resolution images and (2) extending its architecture to incorporate comprehensive multimodal inputs—including image, text, and layout information—specifically optimized for the task of table understanding.

3 Methods

The Table Question Answering (TQA) subtask within the NTCIR-18 U4 is to extract precise answer values from tables in financial reports based on given questions. An optional component of this task involves identifying the specific cell IDs that contain these answers.

We initially attempted to infer cell values using a VLM but found it challenging to estimate exact values, as the answers can differ from the cell contents due to variations in unit descriptions. To address this issue, cell IDs are instead and corresponding cell values are retrieved through rule-based conversion.

To integrate cell IDs into a multimodal model that utilizes both image and layout information, cell IDs are embedded into table images, as illustrated in Figure 1.

3.1 Data Preparation

Figure 2 illustrates the processing flow for obtaining image (I), text (T), and layout (L) modalities from HTML tables with inserted cell IDs. First, the source HTML is segmented into individual tables, and cell IDs are extracted from each cell's attributes and embedded into the HTML content. During this process, a dictionary is generated to map each cell ID to its corresponding value. To convert the HTML tables with inserted cell IDs into image and layout modalities, the tables are rendered as PDFs. From these PDFs, the image, text, and layout modalities (I, T, L) are extracted, where the layout modality consists of bounding-box coordinates that define the position of the text. Finally, an instruction dataset is constructed that provides cell IDs as answers, aligning with the given QA dataset.

3.2 Model Construction

LVLMs typically accepts only images as input. However, for this task, the architecture is modified to incorporate text and layout modalities in addition to images.

Following previous studies such as LayTextLLM [8], layout embedding is achieved by converting bounding box coordinates into the hidden dimensions of an LLM using a two-layer MLP. The layout is input as a single token in the LLM and paired with the text

2020年３月31日 現在
March 31, 2020

r1c1:セグメントの名称 Segment Name	r1c2:従業員数(人) Number of Employees
r2c1:空調・冷凍機事業 Air Conditioning and Refrigeration Business	r2c2:74,466 r3c1:(9,151)
r4c1:化学事業 Chemical Business	r4c2:3,876 r5c1:(264)
r6c1:その他事業 Other Businesses	r6c2:1,077 r7c1:(130)
r8c1:全社(共通) Company-wide (Common)	r8c2:950 r9c1:(43)
r10c1:合計 Total	r10c2:80,369 r11c1:(9,588)

(注) １ 従業員数は就業人員であり、臨時従業員数は( )内に年間の平均人員を外数で記載しております。
The number of employees represents the number of working personnel, and the number of temporary employees is recorded separately in parentheses as an annual average.

Figure 1: Example of cell-id inserted table image

inside the corresponding bounding box, applying this process to all text within the table. This approach enables the model to process text while maintaining its spatial correspondence within the table.

In the experiments, different combinations of image, layout, and text inputs are tested to analyze the contributions of each modality. To ensure consistency in analysis, all evaluations are conducted under the same conditions, allowing for a direct comparison of the impact of each input type.

3.3 Pre-training for Layout Modality

Since the proposed layout modality is not present in existing LVLMs, pre-training is conducted to help the model adapt to this format. The LayoutLLM-SFT dataset [9] is designed for document-based QA tasks and includes OCR text and bounding-box coordinates alongside images and QA data. For experiments involving T+L, the model is pre-trained using 50% of the LayoutLLM-SFT dataset and fine-tuned with the Table QA training data.

3.4 Post-processing of Cell Values

Because the model infers cell IDs, converting cell IDs to cell values is necessary at submission. The dictionary of cell IDs and cell values mentioned above is insufficient as a format for answers. For example, it is necessary to consider converting units such as "million yen". Therefore, we performed rule-based conversion to obtain cell values after inferring cell IDs. The HTML of the entire table contains information about units. Additionally, the question includes information about the required value type (such as amount, number, or date). Using this information along with a dictionary of cell IDs and cell values, a rule-based conversion from cell IDs to cell values is performed.

4 Experiments

In this section, we present the experimental conditions and results using the data and models described in the previous section.

4.1 Training Conditions

Table 1 shows the conditions of the trained models. The method in the table shows that I+T+L is when all image, text, and layout are input, T+L is when only text and layout are input, I+T is when only image and text are input, and I* is when only image

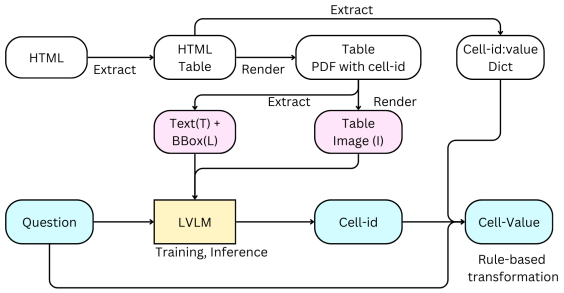

Figure 2: Overall system architecture

Table 1: Main Experiments for training

Method	Description
I+T+L	Training with Image, Text and Layout
T+L	Training with Text and Layout
I+T	Training with Image and Text
I*	Training with Image w/o Pre-Training
markdown (T)	Rendered Markdown from HTML
json (T)	Rendered json from HTML

is input. In addition, training and inference using markdown and json-converted table text was also performed. All experiments are based on LLaVA-Onevision-7B[2, 6], and pre-training using the LayoutLLM-SFT dataset was conducted when including the layout modality. Training for Table QA was conducted under the following conditions: 3 epochs of training, batch size of 8, learning rate of 1e-5, and warmup ratio of 0.03.

4.2 Results

Table 2 presents the comparison results for each modality. Performance discussions refer to the private score, while case studies refer to the public score. The submitted answer corresponds to the I+T+L modality; however, since the image resolution during inference was set lower than expected, it is provided as a reference value.

To compare the zero-shot performance of high-performance multimodal models, we conducted inference using Qwen2.5-VL-72B [12, 13] and GPT-4o (gpt-4o-2024-08-06).

4.2.1 Comparison of Modalities. The highest accuracy was consistently achieved when all modalities—image (I), text (T), and layout (L) were utilized simultaneously. Removal of either the image or layout modality led to a noticeable decrease in performance, especially pronounced when omitting layout. This demonstrates the crucial role that layout information plays in accurately interpreting structured data within tables. The layout modality specifically captures spatial relationships among table elements, providing critical context that is not easily inferable from textual or visual inputs alone.

Further examination of the significance of the layout the layout reveals specific instances (illustrated in Figure 3) where tof the

Table 2: Comparison of Modalities (accuracy)

	Public		Private	
Method	id	value	id	value
submitted	0.9685	0.9163	0.9530	**0.8483**
I+T+L	0.9693	0.9171	**0.9536**	**0.8483**
T+L	**0.9747**	**0.9179**	0.9423	0.8433
I+T	0.9632	0.9140	0.9348	0.8408
I*	0.9087	0.8741	0.8840	0.8006
markdown (T)	0.9586	0.9025	0.9448	0.8382
json (T)	0.9586	0.9010	0.9373	0.8332
Qwen2.5-VL-72B (I)	0.4996	0.4935	0.5003	0.4721
GPT-4o (I)	0.4206	0.4175	0.4188	0.3868

* represents w/o LayoutLLM Pre-training

absence of the layout data rendered the task unsolvable. In this scenario, although explicit (r, c) coordinates were provided within the text, the lack of spatial positioning details made resolving discrepancies between header cell coordinates and content cells infeasible. This highlights that positional context from layout data is essential for precise cell identification. The layout modality effectively bridges the gap between textual and spatial information, enabling accurate contextual understanding of table data.

Performance significantly deteriorated under the image-only condition, particularly for larger tables with smaller textual elements. This underperformance is attributable to limitations inherent in current LVLM OCR methods, which struggle with text recognition at smaller scales or in crowded visual contexts. It shows the necessity of enhanced OCR capabilities or supplementary modalities to reliably parse complex visual data. Additionally, future research might benefit from exploring hybrid approaches that combine image enhancement techniques with multimodal methods to address this limitation.

Moreover, the high performance achieved using standard textual representations (markdown and JSON) suggests the significance of textual modality for structured information extraction. However, the superior results observed in the combined modality (I+T+L) compared to text-only representations emphasize the importance of layout information, showing its effectiveness in capturing generalized structural insights beyond plain textual formats. Thus, integrating comprehensive layout data into multimodal approaches can considerably improve model generalization and robustness.

4.2.2 Performance Comparison with State-of-the-Art Models. We compared the performance of VQA using Qwen2.5-VL-72B and GPT-4o. As with other experiments, we conducted a task to infer cell IDs using table images with inserted cell IDs as input. The modality was conducted under the I-only condition using only images. The zero-shot performance of these models was lower than that of all proposed methods in this study. There was a significant difference compared to the performance when fine-tuned with images only, suggesting that task-specific rules and dataset characteristics are important.

4.2.3 Comparison of Training Data. We investigated the effect of pre-training using the LayoutLLM-SFT dataset when including the Layout modality. Table 3 shows the comparison results of

(a) 2019年度における取締役及び監査役への報酬等の総額は次の通りです。
The total amount of remuneration, etc. for directors and auditors in the fiscal year 2019 is as follows.

r2c1:役員区分 Officer classification	r2c2:報酬等の総額			r2c4:支給人数 Number of recipients
	r3c1:現金報酬 Cash remuneration	r3c2:株式報酬型 Stock compensation ストックオプション	Total amount of remuneration, etc.	
r4c1:取締役 Directors （社外取締役を除く）	r4c2:626百万 million yen 円	r4c3:178百万円	r4c4:804 百万円	r4c5:7 人
r5c1:監査役 Auditors （社外監査役を除く）	r5c2:87百万円	r5c3:-	r5c4:87 百万円	r5c5:3 人
r6c1:社外役員 Outside officers	r6c2:106百万円	r6c3:-	r6c4:106 百万円	r6c5:7 人
r7c1:合計 Total	r7c2:819百万円	r7c3:178百万円	r7c4:997 百万円	r7c5:17 人

(Note 1) Amounts are rounded down to the nearest million yen.
（注）1．金額は、百万円未満を四捨五入しております。

Figure 3: An example table illustrating discrepancies between assigned cell identifiers and the visual layout structure.

Table 3: Comparison of Training Data (accuracy)

	Public		Private	
Method	id	value	id	value
I+T+L	0.9693	0.9171	0.9536	0.8483
I+T+L w/o LayoutLLM	**0.9739**	**0.9179**	**0.9599**	**0.8564**

the presence or absence of pre-training using the LayoutLLM-SFT dataset. When pre-training using the LayoutLLM-SFT dataset was not performed, higher performance was shown in both Public and Private. This finding indicates that the generic LVLM's capability for comprehending structured visual information may inherently be sufficient or even preferable for Table QA tasks, rather than specialized pre-training that could inadvertently constrain model generalization.

5 Conclusions

In this study, we created models using multimodal information such as text, image, and layout for the Table QA task. The characteristics of the modalities showed that the highest accuracy was achieved when all modalities T, I, and L were used. The contributions of the modalities showed that text had the highest contribution, followed by layout and image in order of performance. The findings emphasize that accurate structured data extraction benefits significantly from incorporating comprehensive spatial and textual context, suggesting future directions for enhancing multimodal model architectures in structured data tasks.

References

[1] [n. d.]. https://modelscope.cn/datasets/iic/Layout-Instruction-Data/summary
[2] [n. d.]. https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov
[3] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. *arXiv [cs.CL]* 1 (April 2022).
[4] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A Dataset for Form Understanding in Noisy Scanned Documents. arXiv:1905.13538 [cs.IR] https://arxiv.org/abs/1905.13538
[5] Yasutomo Kimura, Eisaku Sato, Kazuma Kadowaki, and Hokuto Ototake. 2025. Overview of the NTCIR-18 U4 Task. *Proceedings of the 18th NTCIR Conference on Evaluation of Information Access Technologies* (6 2025).
[6] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *arXiv [cs.CV]* (Aug. 2024).
[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *arXiv [cs.CV]* (April 2023).
[8] Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, Hao Liu, and Can Huang. 2024. A bounding box is worth one token: Interleaving Layout and text in a Large Language Model for document understanding. *arXiv [cs.CL]* (July 2024).
[9] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15630–15640.
[10] Minesh Mathew, Dimosthenis Karatzas, and C V Jawahar. 2020. DocVQA: A Dataset for VQA on Document Images. *arXiv [cs.CV]* (July 2020).
[11] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. [n. d.]. CORD: A Consolidated Receipt Dataset for Post-OCR Parsing. ([n. d.]).
[12] Qwen Team. 2025. Qwen2.5-VL. https://qwenlm.github.io/blog/qwen2.5-vl/
[13] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv [cs.CV]* (Sept. 2024).
[14] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2019. LayoutLM: Pre-training of text and layout for document image understanding. *arXiv [cs.CL]* (Dec. 2019).
[15] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2020. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. *arXiv [cs.CL]* (Dec. 2020).