STMK24 NTCIR18 U4 Table QA Submission

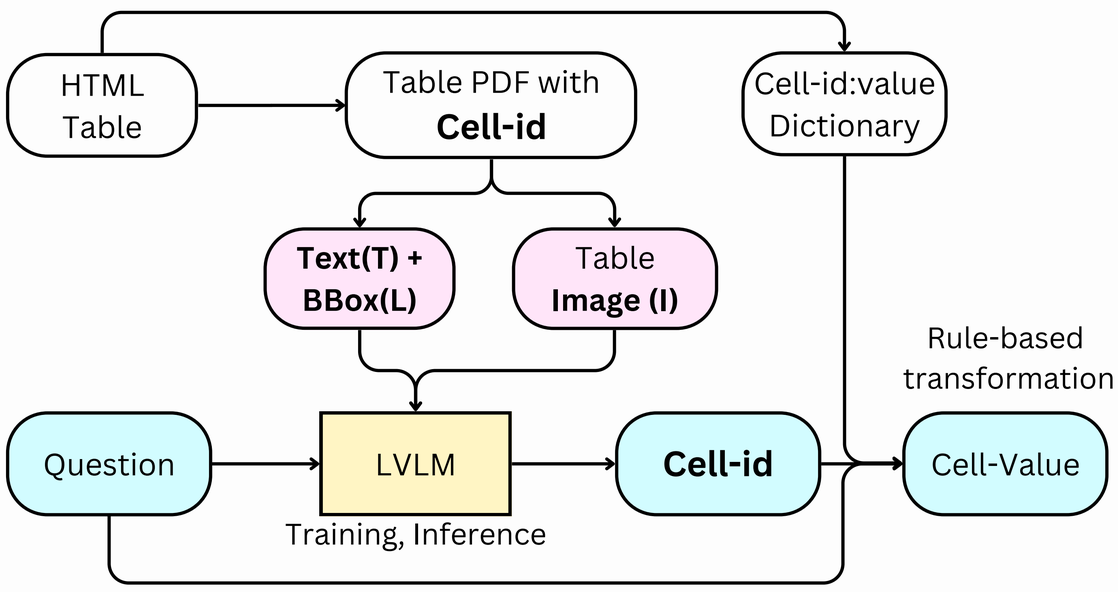
Hayato Aida, Kosuke Takahashi, Takahiro Omi
Stockmark, Japan

Introduction

- Tables in business documents appear in diverse formats, complicating processing.
- NTCIR-18 U4 provides HTML tables from Japanese reports, requiring answer extraction.
- With recent Large Vision-Language model (LVLM) advances, we explore table understanding via image, text, and layout.
- Our model predicts answer cell IDs and uses rules to extract values.

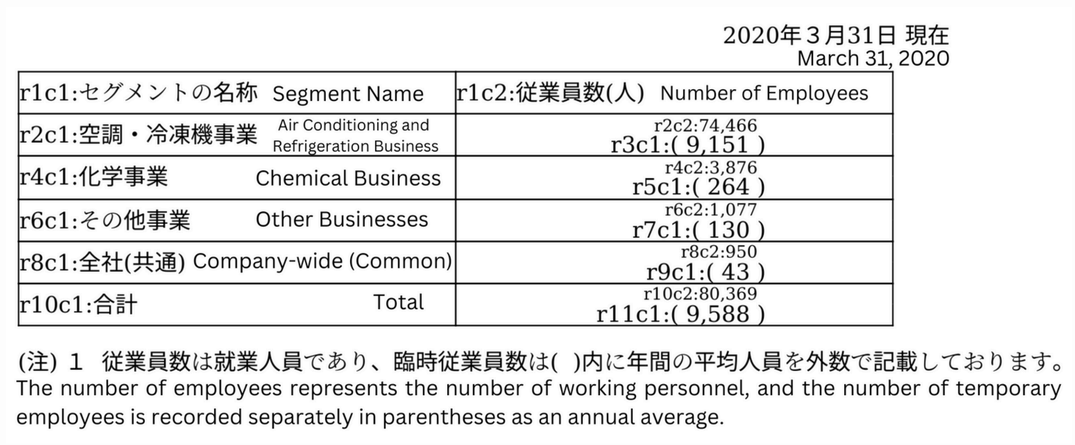
Key Points of Our Method

- Table images with embedded **Cell-id**
- Multimodal table QA leveraging **image, layout, and text** information extracted from tables



Cell-id Embedding

- Cell IDs are inserted into each HTML table cell and rendered into a PDF and image.
- This makes each cell visually identifiable by the model without relying on structural tags.
- A dictionary mapping Cell-id to value is created during preprocessing to convert predicted IDs into answer values.



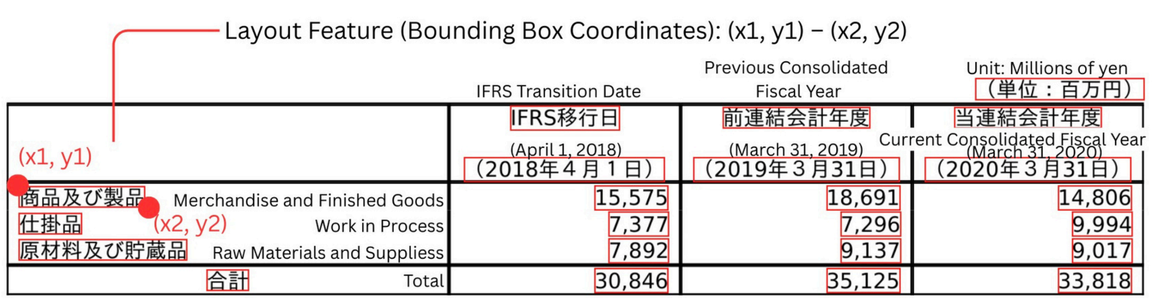
2020年３月31日 現在
March 31, 2020

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.

Layout and Text Modality

- Layout refers to bounding-box coordinates surrounding each text block.
- These coordinates are projected into the model's hidden space via a 2-layer MLP.
- The model can capture relationships between table elements such as headers and contents.

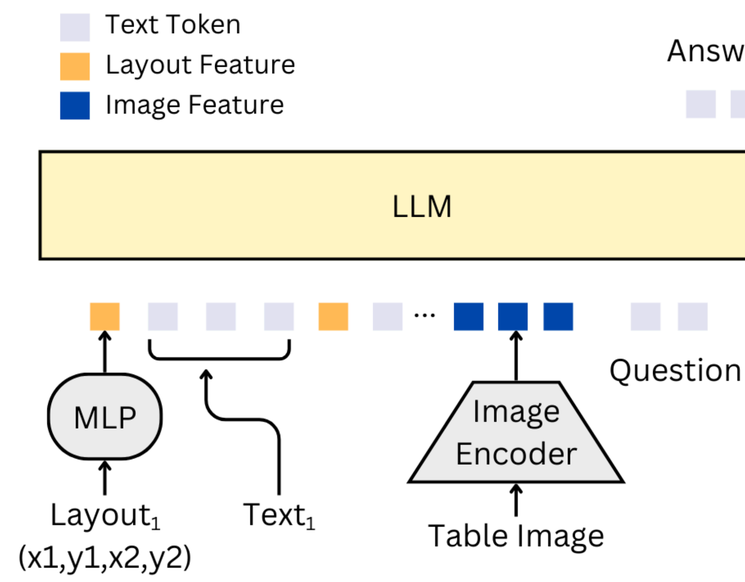


	IFRS Transition Date	Previous Consolidated Fiscal Year	Current Consolidated Fiscal Year
	IFRS移行日 (April 1, 2018) (2018年４月１日)	前連結会計年度 (March 31, 2019) (2019年３月31日)	当連結会計年度 (March 31, 2020) (2020年３月31日)
商品及び製品 Merchandise and Finished Goods	15,575	18,691	14,806
仕掛品 Work in Process	7,377	7,296	9,994
原材料及び貯蔵品 Raw Materials and Suppliess	7,892	9,137	9,017
合計 Total	30,846	35,125	33,818

Unit: Millions of yen (単位：百万円)

Layout-Aware LVLM Architecture

Each text token is paired with its bounding-box layout, encoded via MLP and fused with image features as LLM input.



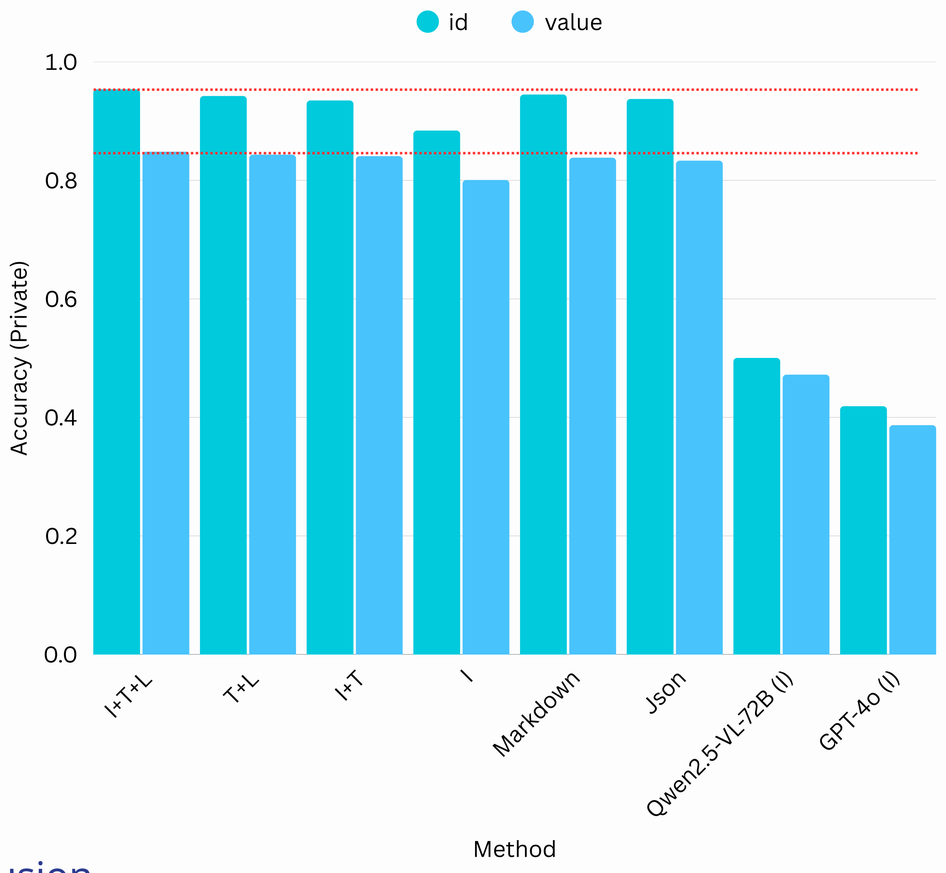
Experiments

- Base model: LLaVA-OneVision-7B.
- Pre-training for layout modality used 50% of LayoutLLM-SFT, followed by fine-tuning on Table QA.
- Also benchmarked zero-shot performance of Qwen2.5-VL-72B and GPT-4o (I-only condition).

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**

Results

- I+T+L consistently achieved the highest accuracy (95.36% ID accuracy on private score).
- Excluding layout or image reduced performance, with layout contributing more than image.
- Image-only (I*) performance degraded on small-text tables, revealing OCR limitations in current LVLMs.
- Markdown and JSON text formats yielded strong performance but fell short of full multimodal results.



Conclusion

- Developed a multimodal Table QA model **integrating image, text, and layout features**.
- Found that layout plays a pivotal role, second only to text in modality importance.
- These results point to promising future directions in structural-aware multimodal LLM architectures.