Language Technology

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

EMNLP 2025
Suzhou, China | 中国苏州

MTabVQA: Evaluating Multi-Tabular Reasoning of Language Models in Visual Space

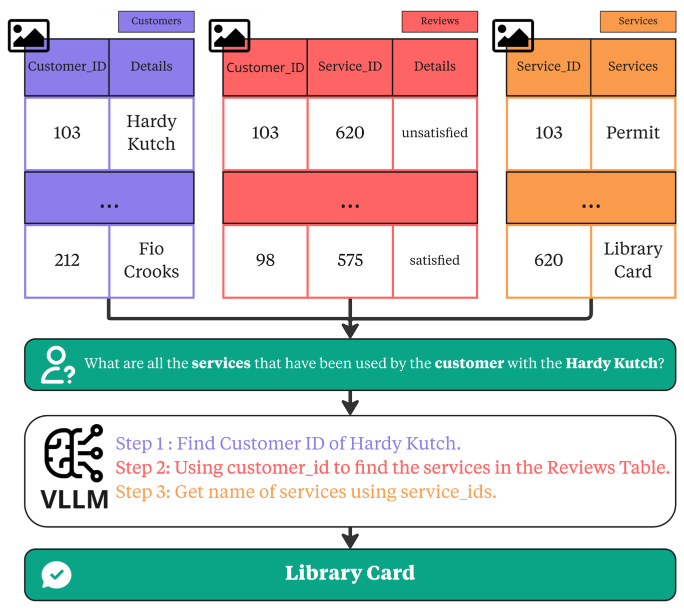
Anshul Singh[1], Chris Biemann[2] & Jan Strich[2]

[1] Department of Information Technology, Panjab University | [2] Language Technology Group, Universität Hamburg

🤔 Why Do We Need this Benchmark?

Vision-Language Models (VLMs) excel at layout understanding, but fail when reasoning requires synthesizing information from multiple, visually rendered tables a common real-world task.

- **Critical Gap**: Existing benchmarks are text-based or focus on single tables.
- **Real-World Need**: Web pages, reports, and scanned documents present data visually across multiple tables.
- **Complex Task**: Requires robust OCR, layout parsing, cross-table entity linking, and multi-hop logical reasoning, all from pixels.



F1 Figure 1: MTabVQA requires models to link Hardy Kutch to Customer_ID 103, then use that ID to find Service_ID 620 in a second table, and finally look up the service name Library Card in a third table.

The MTabVQA Benchmark

We introduce **MTabVQA**, a comprehensive benchmark suite designed to bridge this gap.

- **MTabVQA Benchmark**
  - 3,745 complex question-answer pairs.
  - Requires reasoning across 2 to 5 separate table images per question.
  - Covers 14 distinct reasoning categories (e.g., aggregation, comparison, fact-checking).
- **MTabVQA-Instruct**
  - A large-scale instruction tuning dataset with 15,853 examples to enhance VLM capabilities.

Dataset Split	Source	Sub-dataset	#QA Pairs	#Tables	Proportion (%)
MTabVQA	QFMTS (Zhang et al., 2024b)	MTabVQA-Query	2456	5541	65.7%
	Spider (Yu et al., 2018)	MTabVQA-Spider	1048	2363	27.9%
	Atis (Dahl et al., 1994)	MTabVQA-Atis	112	429	3.0%
	MiMoTable (Li et al., 2025b)	MTabVQA-Mimo	129	166	3.4%
	Total Eval Set		**3745**	**8499**	100.0%
MTabVQA-Instruct	MultiTabQA (Pal et al., 2023)	–	10,990	21,976	69.3%
	Spider (Yu et al., 2018)	–	2395	5845	15.2%
	BIRD (Li et al., 2023a)	–	1572	3144	9.9%
	Atis (Dahl et al., 1994)	–	384	1780	2.4%
	MiMoTable (Li et al., 2025b)	–	512	719	3.2%
	Full Instruct Set		**15,853**	**33,464**	100.0%

T1 Table 1: Detailed composition of the MTabVQA and MTabVQA-Instruct datasets. The table shows the original data sources and provides statistics for each sub-dataset, including the number of QA pairs and unique tables.

Key Features:

Visual-First: Tables are rendered as diverse images, simulating real-world appearance with 10+ unique visual themes.

Multi-Hop Necessity: Questions are constructed to be unanswerable without correlating data across multiple tables.

🛠️ Dataset Construction Framework

We developed a framework to generate high-quality, visually-grounded question-answer pairs that necessitate multi-table reasoning.

Data Sourcing & Relational Sampling

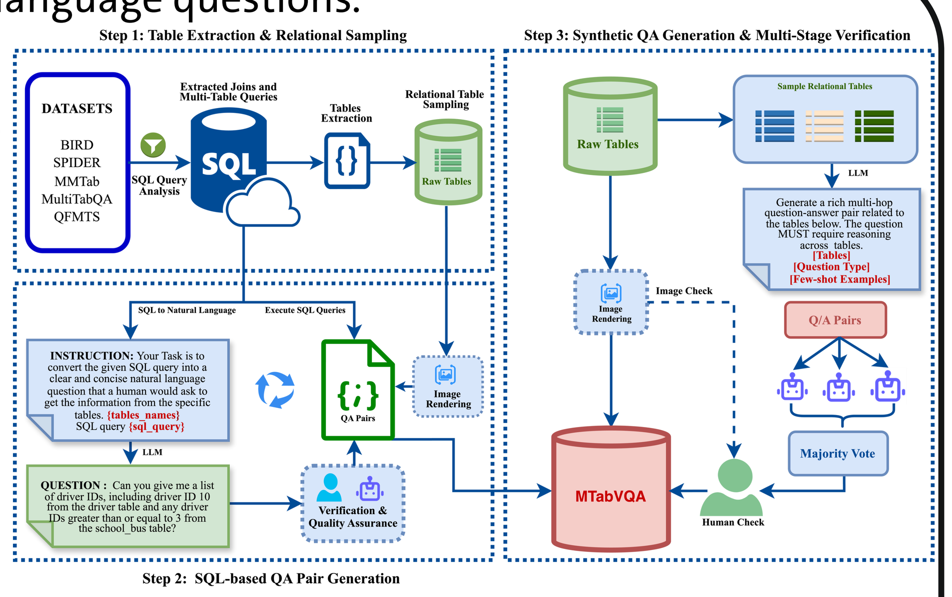
- Extracted multi-join queries from 6 diverse datasets.
- Used a graph-based sampling algorithm to create smaller, interconnected table subsets while preserving relational integrity.

Multi-Hop QA Generation

- Converted complex SQL queries into natural language questions.
- LLMs to generate QA pairs based on a predefined taxonomy of 14 reasoning types (e.g., aggregation, ranking, fact-checking).

Rendering & Verification

- Visually Diverse Rendering of tables into images.
- Multi-LLM agent system for automated checks, followed by final human verification.



F2 Figure 2: MTabVQA Construction Framework Overview. (1) Data Sourcing & Sampling: Identify multi-table relational data via SQL joins, extract tables, apply relational sampling. (2) Visual QA Generation: Generate multi-hop QA pairs via SQL-to-question conversion or LLM-guided generation from sampled tables/taxonomy; render tables as images. (3) Verification & Finalization: Apply automated (LLM) and human verification for quality and multi-table necessity.

📊 How Do VLMS Perform?

We evaluated leading open-source and proprietary VLMs in a zero-shot setting.

- **Key Takeaway 1:** Open-source models perform poorly out-of-the-box.
- **Key Takeaway 2:** Even powerful proprietary models like GPT-4.1 are far from perfect.
- **Key Takeaway 3:** Our fine-tuned TableVision outperforms all models, demonstrating the effectiveness of targeted instruction tuning with MTabVQA-Instruct.

❗ *Visual multi-tabular reasoning remains a significant challenge.*

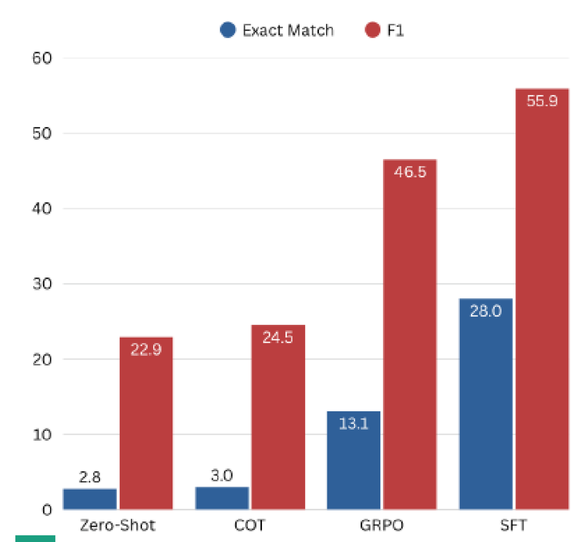
Model	MTabVQA-Spider				MTabVQA-Query				MTabVQA-ATIS				MTabVQA-MiMo				Overall	
	EM	F1	P	R	EM	F1	P	R	EM	F1	P	R	EM	F1	P	R	EM	F1
Open-Source VLMs (Zero-Shot)																		
LLaVA-OV-Qwen2-7B	2.2	20.0	19.5	29.3	2.3	15.7	15.9	23.6	0.0	9.2	5.9	33.8	0.7	5.5	4.3	19.1	2.1	18.4
Phi-3.5-Vision-Instruct	2.9	26.1	25.9	39.6	2.4	22.0	22.3	34.7	1.8	15.0	15.3	24.8	0.8	3.2	3.6	3.3	2.5	22.3
InternVL3-8B-Instruct	6.1	32.4	33.0	39.1	5.2	24.8	26.9	29.6	3.6	20.3	19.5	31.9	7.0	19.1	22.3	21.3	5.4	26.6
Qwen2.5-VL-7B	8.0	39.8	40.4	44.0	7.8	33.9	34.8	38.0	6.3	32.6	29.0	48.6	9.3	22.2	25.9	22.8	7.8	35.1
Gemma-3-12B-IT	15.6	48.0	48.2	53.4	10.3	38.1	39.4	42.6	11.6	35.1	34.2	40.8	9.3	18.6	22.0	18.8	**11.8**	**40.1**
Proprietary VLMs (Zero-Shot)																		
Gemini-2.0-Flash	42.9	68.5	69.2	71.2	31.4	57.3	58.2	60.5	22.3	36.0	37.2	37.5	24.0	42.3	49.2	41.2	34.1	59.3
GPT-4.1	49.0	74.3	74.7	76.6	34.2	58.5	59.2	60.8	6.3	39.9	30.0	86.3	20.2	39.6	44.9	38.8	**37.0**	**61.7**
Fine-tuned Model (Ours)																		
TableVision (Ours)	32.4	64.3	66.6	66.1	49.8	72.6	74.0	73.5	33.0	45.9	48.4	47.8	20.1	36.2	40.8	36.4	**43.4**	**68.2**

T2 Table 2: Performance Comparison of VLMs on MTabVQA Sub-datasets (%), and Overall EM/F1 (%). Models categorized and sorted by overall F1 score within categories. Overall scores are weighted averages. Best overall and best open-source zero-shot overall scores are bolded. EM denotes Exact Match, P Precision, and R Recall.

💡 Key Insights & Analysis

Fine-Tuning is the Most Effective Strategy

Supervised Fine-Tuning (SFT) on our MTabVQA-Instruct dataset provides massive gains, significantly outperforming advanced prompting (CoT) and reinforcement learning (GRPO) techniques.



F3 Figure 3: Performance comparison of Qwen2.5-VL-3B on the MTabVQA with different post-training strategies.

Data Diversity Trumps Scale for Generalization

Training on our full, diverse dataset (TableVision) yielded the best overall performance. A model trained on a larger but narrowly-focused dataset (MultiTabQA subset) generalized poorly, showing that exposure to varied table structures and reasoning types is crucial.

Fine-tuning Subset (Source)	# Samples	Overall	
		EM	F1
Qwen2.5-VL-7B (Zero-Shot)	0	7.8	35.1
MiMo+ATIS Subset	896	13.0	40.0
Spider Subset	2,395	41.5	65.2
MultiTabQA Subset	10,990	9.4	30.2
MTabVQA-Instruct (Full)	15,853	**43.4**	**68.2**

T3 Table 3: Performance of fine-tuned models on dataset splits of MTabVQA-Instruct, measuring the influence of the dataset on the overall performance on MTabVQA. Performance is measured in EM and F1

🎬 Conclusion

- We introduced MTabVQA, a new benchmark to evaluate and advance multi-tabular reasoning in the visual domain.
- *Our results reveal significant limitations in current SOTA VLMs and show that our fine-tuned TableVision model sets a new performance benchmark.*
- **Future Directions:** Expanding to more complex layouts (merged cells, embedded charts), non-English tables, and programmatically-aided reasoning.

Scan for Code and Dataset
anshulsc.live/MTabVQA

hcds.uni-hamburg.de