TABVERSE: Benchmarking Cross-Format Table Understanding in LLMs and VLMs

Anonymous ACL submission

Abstract

Tables are a fundamental medium for presenting information in domains such as finance, healthcare, and education, making table comprehension a critical capability for artificial intelligence models and agents. While recent benchmarks have focused on question-answering performance on rendered table images, two important questions remain unexplored: (i) To what extent do models truly understand rendered table images? and (ii) How does performance vary across different structural formats, such as HTML, Markdown, and LaTeX? To address these questions, we introduce TABVERSE, a dataset benchmark that analyzes tables in various structural formats, including HTML, Markdown, and LaTeX, along with their corresponding rendered visualizations. The benchmark evaluates model performance across three tasks: table-based question answering, structural understanding, and structure reconstruction. The evaluation results reveals consistent advantages for vision-language models over large language models across all tasks, while span-based question answering and structure reconstruction still pose significant challenges. These findings underscore the importance of visual grounding and expose persistent limitations in information localization and structural information recovery. Data and code are available at `http://URL.withheld.for.review`.

1 Introduction

Tables present information in a compact and structured format that facilitates human comprehension and analysis, making them a fundamental medium in financial reports, laboratory studies, and research papers (Barnabò, 2023; Smock et al., 2022). Unlike plain text, tables use structural elements such as rows, columns, and headers to organize information and convey meaning. As Artificial Intelligence (AI) systems increasingly rely on Large Language Models (LLMs) and Vision-Language Models (VLMs), it becomes essential for these models to develop the capability to understand tabular information (Liu et al., 2022; Sui et al., 2024a). This raises important questions: *Can LLMs and VLMs effectively comprehend tables, and to what extent?*

Tables are commonly represented on the Internet in three structural formats, but are not limited to: **HTML** for webpages, **Markdown** for code repositories such as GitHub, and **LaTeX** for academic and scientific publications. While all of these formats present data in a tabular structure, they often differ significantly in syntax and styling. HTML supports nested structures and rich styling, Markdown emphasizes simplicity and minimalist plain text-driven design with limited formatting options, and LaTeX prioritizes typographic precision frequently involving complex commands.

These stylistic and structural discrepancies create major challenges for AI systems in table understanding. First, format-specific syntax introduces parsing difficulties when models are not trained to distinguish between HTML tags, Markdown pipes, or LaTeX commands. Second, identical content can appear in different structural or visual forms, requiring separation of presentation from meaning. Third, cues such as merged cells, multi-level headers, or custom alignments are often lost in plain text, demanding reasoning beyond surface patterns. Overall, the heterogeneity of table formats and renderings hinders the development of generalizable models for cross-format and multimodal table comprehension.

To evaluate the table comprehension ability of LLMs and VLMs, we construct and release the TABVERSE benchmark. TABVERSE combines and extends records from five widely used datasets, namely FEVEROUS (Aly et al., 2021), HYBRIDQA (Chen et al., 2020b), SQA (Iyyer et al., 2017), TABFACT (Chen et al., 2020a), and TOTTO (Parikh et al., 2020), to ensure diverse table content coverage.

In addition to the original tasks in these datasets, TABVERSE introduces annotations for two complementary tasks: *Structural Understanding Capability (SUC)* and *Structure Reconstruction (SR)*. SUC, evaluates a model's ability to reason over fine-grained table structures by performing localized sub-tasks. The latter involves (*i*) recovering structural formats (HTML, Markdown, or LaTeX) from table images and (*ii*) translating across textual formats using VLM conditioned on visual representations.

Our work makes the following contributions:

- **A benchmark for cross-representation analysis.** We release over 10,000 tables aligned across HTML, Markdown, LaTeX, and images, enabling direct comparison of LLMs and VLMs on identical content without representation bias.
- **A task suite covering core table abilities.** The benchmark includes SUC, question answering, and SR, allowing clear separation of structural and reasoning skills.
- **An evaluation setup isolating visual contribution.** Pairing each VLM with its base LLM quantifies the specific value added by visual input versus text alone.
- **Insights for future multimodal benchmarks.** Our findings reveal systematic effects of format and recurring reconstruction errors, offering guidance for dataset and model design.

2 Related Work

Research on table reasoning has advanced along two main directions: **text-based** and **vision-based** approaches. Early benchmarks such as SQA (Iyyer et al., 2017), TABFACT (Chen et al., 2020a), TOTTO (Parikh et al., 2020), and HYBRIDQA (Chen et al., 2020b) introduced tasks like sequential QA, fact verification, and table-to-text generation, but relied solely on serialized text. Subsequent work focused on structural reasoning. Table Meets LLM (Sui et al., 2024a) examined partitioning and retrieval, revealing strong dependence on serialization, while TAP4LLM (Sui et al., 2024b) explored sampling and augmentation for large tables. Robustness studies (Singha et al., 2023) evaluated model stability under perturbations. More recent efforts, including TableEval (Zhu et al., 2025) and RealHiTBench (Wu et al., 2025), compared formats such as HTML, Markdown, and LaTeX and introduced hierarchical or multimodal renderings, yet most research remains confined to text-based inputs.

In parallel, **visual table reasoning** has been explored. TableVQA (Kim et al., 2024) rendered tables as images for QA and found that vision language models perform worse than text models. MTabVQA (Singh et al., 2025) extended this to multiple languages, TabPedia (Zhao et al., 2024) combined detection, structure recognition, and QA, and Image2Struct (Roberts et al., 2025) reconstructed HTML or LaTeX from images. These efforts advanced visual reasoning but did not directly compare vision and text on the same data. Recent work such as ALTER (Zhang et al., 2025) improved large table reasoning by selecting key subsets and adding schema and semantic context, which increased scalability and robustness for language models.

Overall, previous studies explored different forms of tables but not in a single controlled setup. Few have compared vision models with their text counterparts, so the benefit of visual input remains unclear. TABVERSE fills this gap by providing tables in HTML, Markdown, LaTeX, and image forms, and by evaluating *LLM only*, *VLM only*, and *Text-only VLM* models on the same content for structure understanding, QA, and reconstruction.

3 Dataset

We introduce TABVERSE, a comprehensive dataset built from samples of five established table question-answering benchmarks: FEVEROUS (Aly et al., 2021), HYBRIDQA (Chen et al., 2020b), SQA (Iyyer et al., 2017), TABFACT (Chen et al., 2020a), and TOTTO (Parikh et al., 2020). The dataset represents each record in three structural formats, HTML, Markdown, and LaTeX, and renders them into visual images for performance evaluation. These multiple formats and renderings provide a controlled setting to systematically evaluate and analyze how format and modality affect a model's ability to understand tables, while supporting direct comparisons between text-only and vision-based reasoning.

3.1 Source Datasets

TABVERSE contains 5,000 query–table pairs, with 1,000 pairs sampled from each table benchmark.

Dataset	QA Pairs	Unique Tables	JSON	HTML	Markdown	LaTeX	Images	Alignment
FEVEROUS	1,000	635	✓	✗	✗	✗	✗	✗
HYBRIDQA	1,000	967	✓	✗	✗	✗	✗	✗
SQA	1,000	148	✓	✗	✗	✗	✗	✗
TABFACT	1,000	740	✓	✗	✗	✗	✗	✗
TOTTO	1,000	995	✓	✗	✗	✗	✗	✗
Table Meets LLM (Sui et al., 2024a)	5000	3,485	✓	✓	✓	✗	✗	✗
Tap4LLM (Sui et al., 2024b)	36,253	13,514	✓	✓	✓	✗	✗	✗
TableVQA (Kim et al., 2024)	1500	894	✗	✓	✗	✗	✓	✗
TABVERSE	**5,000**	**3,485**	✓	✓	✓	✓	✓	✓

Table 1: Comparison of five source datasets and benchmarks with **TABVERSE**. Previous datasets mostly offer tables in JSON or limited formats without visual renderings or cross-format alignment. TABVERSE unifies 5,000 QA pairs and 3,485 tables across HTML, Markdown, and LaTeX, each paired with rendered images, enabling fully aligned multimodal and multi-format table reasoning.

This sampling strategy ensures that TABVERSE covers four diverse domains: fact verification (FEVEROUS, TABFACT), multi-hop question answering (HYBRIDQA), sequential question answering (SQA), and table-to-text generation (TOTTO). Our analysis shows that TABVERSE contains 3,485 unique tables, as certain tables are associated with multiple queries. Table 1 provides a comparison between the five source datasets and related benchmarks with **TABVERSE**, highlighting its broader multimodal and multi-format coverage.

3.2 Format Conversion

Each unique table is converted into three structural formats, HTML, Markdown, and LaTeX, using a linearization method adapted from Sui et al. (2024a). Table metadata are serialized into the target syntax while preserving layout and content consistency. Specifically, HTML uses `<table>`, `<tr>`, `<th>`, and `<td>` tags; Markdown employs pipe-delimited tables via `tabulate`[1]; and LaTeX generates compilable tabular code through `pandas`[2] with proper escaping of special characters. All conversions use `numpy`[3] and `pandas` for consistent alignment, producing 10,455 semantically equivalent tables across formats.

3.3 Rendering Pipelines

Each of the 10,455 structural formats is rendered as an image using format-specific pipelines to support visual comparison. These include: (i) an HTML renderer using Playwright[4] for browser-based rendering and CSS styling, (ii) a Markdown renderer that converts Markdown to styled HTML using `markdown2`[5] and captures screenshots with Selenium[6], and (iii) a LaTeX renderer that compiles to PDF via `latexmk`[7] or `pdflatex`[8], then converts the output to PNG using `pdftoppm`[9]. Figure 1 shows the full pipeline from format conversion to rendering.

3.4 Evaluation Configurations

TABVERSE evaluates a model's ability to understand tables through three tasks: question answering, structural understanding, and structure reconstruction. The *question-answering* task assesses how well the model handles real-world tasks, while the *structural understanding* and *structure reconstruction* tasks evaluate the model's ability to interpret and reproduce the table layout independently of its content respectively.

3.4.1 Question Answering

The question-answering task evaluates a model's ability to reason over tabular data and extract relevant information for a given query. We use question–table pairs from FEVEROUS, HYBRIDQA, SQA, TABFACT, and TOTTO to ensure consistency and comparability with existing benchmarks. This task assesses whether models can perform complex reasoning such as multi-hop retrieval, fact verification, and sequential inference. Example prompts are provided in Table 8 of Appendix A.

[1] https://pypi.org/project/tabulate/
[2] https://pandas.pydata.org/
[3] https://numpy.org/
[4] https://playwright.dev/python/
[5] https://github.com/trentm/python-markdown2
[6] https://www.selenium.dev/documentation/webdriver/
[7] https://ctan.org/pkg/latexmk
[8] https://ctan.org/pkg/pdftex
[9] https://www.freedesktop.org/wiki/Software/poppler/

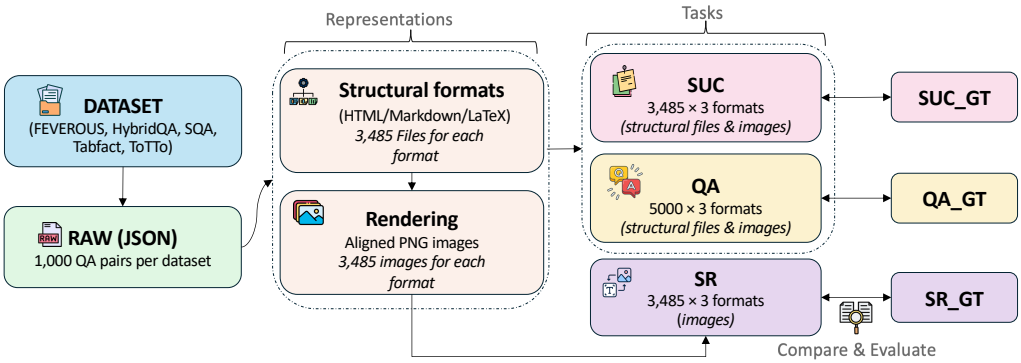

Figure 1: Overview of **TabVerse**. Each table is represented in three structural formats (HTML, Markdown, LaTeX) with corresponding rendered images. These aligned multimodal pairs enable evaluation on QA, SUC, and SR tasks for consistent cross-format and cross-modality analysis.

3.4.2 Structural Understanding

The structural understanding task evaluates a model's ability to detect and interpret table structure through ten subtasks in four categories: (*i*) Partitioning and boundary detection (e.g., identify first cell value); (*ii*) Size estimation (e.g., the number of row and column); (*iii*) Value lookup (e.g., find a cell by value); (*iv*) Row and column retrieval. By asking the same questions across different structural formats, we examine whether format differences affect how models locate and reason over relevant table content. We extend the Structural Understanding Capabilities benchmark Sui et al. (2024a) to include aspects relevant to document parsing and structured data understanding. Example task definitions and model prompts are provided in Table 7 of Appendix A, and the overall task taxonomy is illustrated in Figure 3 of Appendix B.

3.4.3 Structure Reconstruction

The structure reconstruction task evaluates a vision-language model's ability to reconstruct the table structure within the rendered images. Given a table image, the model is prompted to generate the table structure x' in one of three possible formats: HTML, Markdown, or LaTeX. The generated output is then compared with the ground-truth x, after normalizing whitespace and syntax. We report `GriTS` topology and content scores (Smock et al., 2023), along with positional accuracies at the cell, column and row levels to assess structural fidelity and alignment quality.

4 Experimental Settings

This paper examines how table representation and model modality affect table reasoning and comprehension performance. The designed experiments support two important questions: (*i*) How sensitive the models are to structural representation and (*ii*) How much additional information visual input provides beyond text. These questions guide the design of the evaluation pipelines and the analysis of the experiment.

4.1 Experimental Design

We design three complementary evaluation pipelines to study the effect of modality and representation while keeping content identical across settings. (i) *VLM pipeline* evaluates VLMs directly on rendered table images to test if visual encoders capture spatial cues such as layout and cell positioning that text may lose. (ii) *Text-only VLM pipeline* disables the visual encoder of the same models to isolate the role of the language backbone. (iii) *LLM pipeline* evaluates language-only models on the three structural formats to measure their ability to interpret structural syntax.

These pipelines enable controlled comparisons: VLM vs. Text-only VLM to isolate vision effects, and VLM vs. LLM to quantify visual input benefits. Figure 2 in Appendix B shows the overall workflow. This design also examines whether visual information is necessary for structure-based reasoning or if text alone is sufficient.

4.2 Implementation

We use a uniform zero-shot setup for all tasks. Models are evaluated without fine-tuning and generate outputs using greedy decoding (`temperature=0`, `top-p=1`) for deterministic results. Output length is adjusted by task: short for SUC, medium for QA, and long for reconstruction. Light post-processing normalizes outputs and formatting. This includes removing phrases such as *"The answer is"*, convert-

Models	Formats	Detection			Size Estimation			Cell Access			
		T.P.	F.C.	L.C.	S.D.	# Rows	# Cols	C.L.	R.L.	C.R.	R.R.
S-VLM	HTML	0.0	29.7	24.7	**1.4**	**26.5**	**36.9**	**1.8**	**18.4**	**31.7**	0.0
	LaTeX	0.0	23.7	17.8	0.5	15.8	11.2	0.7	14.2	26.1	0.0
	Markdown	0.0	**33.7**	**30.9**	0.3	15.7	20.0	1.1	16.9	29.3	0.0
Q-2.5 3B	HTML	3.6	**70.5**	45.9	33.8	**61.7**	23.7	9.1	24.8	41.5	**15.4**
	LaTeX	11.2	61.1	45.5	**41.7**	49.2	**75.4**	**11.4**	**29.2**	**76.8**	8.8
	Markdown	**18.5**	45.1	**54.1**	38.2	54.0	56.9	4.0	22.5	65.5	14.7
Q-2.5 7B	HTML	8.5	**72.2**	44.3	10.2	**80.9**	26.9	15.9	20.6	30.4	**20.1**
	LaTeX	**15.8**	64.5	45.5	**83.1**	70.7	**90.1**	**18.7**	**21.6**	**68.3**	17.7
	Markdown	12.8	61.6	**55.7**	52.1	55.1	79.9	17.2	18.0	52.2	17.6
GPT-5	HTML	**56.3**	**57.8**	**75.4**	39.7	**96.8**	65.3	41.2	52.1	69.0	**68.9**
	LaTeX	36.2	54.9	70.8	**60.3**	78.6	**83.1**	**43.3**	46.9	**86.6**	44.6
	Markdown	43.1	55.9	73.2	57.5	91.2	79.5	42.3	**53.7**	84.4	52.7

Table 2: Performance of VLMs on the SUC benchmark for the HYBRIDQA dataset across three table formats (HTML, LaTeX, Markdown). Models: **S-VLM** (SMOLVLM-1.7B), **Q-2.5 3B/7B** (QWEN2.5-VL variants), and **GPT-5**. Metrics cover *Partition and Boundary Detection* (T.P.=Table Partition, F.C.=First Cell, L.C.=Last Cell), *Size Estimation* (S.D.=Size Detection, Rows, Cols), and *Cell Access* (C.L.=Cell Lookup, R.L.=Reverse Lookup, C.R.=Column Retrieval, R.R.=Row Retrieval). Full SUC results appear in Table 10, Appendix C.

ing binary responses to 0/1, and applying regular expressions for consistent formatting. These steps ensure fair comparison across models and tasks.

4.3 Models Evaluated

We evaluate open-weight models for transparency and reproducibility. Each VLM is paired with its base LLM to isolate the contribution of visual processing. The evaluated models include SMOLVLM-1.7B (Marafioti et al., 2025) with its base SMOLLM2 (Allal et al., 2025), QWEN2.5-VL (Bai et al., 2025) in 3B and 7B variants with corresponding QWEN2.5 (Yang et al., 2024) base models, and GPT-5 (OpenAI, 2025). Unlike the others, GPT-5 is a closed-source model accessed through the OpenAI API, included here for reference as a high-performance proprietary baseline. This setup enables direct comparison between multimodal and text-only configurations under identical evaluation conditions.

5 Results

We report results across the three evaluation pipelines, *VLM*, *Text-only VLM*, and *LLM*, covering SUC, QA, and SR tasks. This section analyzes how table representation and modality affect performance. Unless stated otherwise, SUC and QA use exact-match accuracy, while structure reconstruction uses `GriTS` (Smock et al., 2023) and positional accuracies for layout and cell alignment.

5.1 SUC

The SUC task tests a model's ability to recover table structure and content across modalities. Table 2 highlights typical outcomes within the VLM pipeline for HYBRIDQA dataset; Full results are provided in Tables 10–12 (Appendix C).

VLM Pipeline: Models process table *images* rendered from HTML, LaTeX, and Markdown. Performance follows a clear hierarchy: GPT-5 > QWEN2.5-VL 7B > QWEN2.5-VL 3B > SMOLVLM-1.7B. Scaling from 3B to 7B improves detection and size estimation, while GPT-5 achieves the best overall accuracy. On HYBRIDQA, VLMs excel at size estimation and column retrieval for LaTeX and Markdown, where regular alignment provides strong spatial cues. Row retrieval is weaker, as it relies more on semantics than geometry, and smaller models underperform on first- and last-cell detection, indicating limited localized grounding.

Across formats, LaTeX and Markdown favor structure-based subtasks (e.g., column counting), while HTML performs better on row-oriented ones due to sequential layout. Among datasets, TABFACT achieves the highest scores, followed by FEVEROUS; SQA performs moderately, and TOTTO remains the most challenging due to its need for both structural and semantic reasoning.

Dataset	Format	VLM				VLM+Text			LLM			
		S-VLM	Q-2.5 3B	Q-2.5 7B	GPT-5	S-VLM	Q-2.5 3B	Q-2.5 7B	S-VLM	Q-2.5 3B	Q-2.5 7B	GPT-5
FEVEROUS	HTML	57.5	**71.2**	75.3	**77.6**	51.8	**64.3**	69.7	**53.0**	53.1	65.9	79.4
	Markdown	58.6	70.6	**77.2**	77.3	**56.5**	62.5	**70.8**	50.3	**56.6**	67.5	**80.4**
	LaTeX	**59.3**	71.1	76.6	76.4	55.9	61.1	69.2	43.8	50.8	**67.6**	79.0
HybridQA	HTML	6.1	10.7	12.3	26.0	2.8	8.2	13.2	2.1	8.3	10.6	32.3
	Markdown	17.4	25.4	34.5	**40.6**	8.7	**29.3**	**35.4**	**7.0**	25.5	33.0	46.2
	LaTeX	**19.1**	**28.9**	**34.8**	38.9	**8.8**	28.9	35.3	6.8	**26.7**	**34.2**	**46.7**
TabFact	HTML	55.0	**71.2**	74.9	77.8	50.0	**61.5**	63.2	50.9	**54.4**	**64.2**	**82.5**
	Markdown	54.6	69.5	74.7	78.2	50.5	60.0	**65.7**	**52.4**	52.9	62.8	79.3
	LaTeX	**56.2**	71.1	**75.5**	**77.7**	**50.8**	57.2	64.5	48.7	52.3	59.6	80.2
SQA	HTML	21.3	**32.7**	**33.7**	30.3	16.8	29.7	**32.4**	12.2	26.9	**29.9**	**33.4**
	Markdown	**22.3**	32.2	33.3	31.0	**19.0**	**30.3**	31.4	**13.8**	**30.9**	28.8	**32.6**
	LaTeX	21.9	30.5	31.9	29.6	16.5	28.6	30.2	8.5	24.8	28.1	**32.6**

Table 3: Accuracy for QA tasks on FEVEROUS, HYBRIDQA, TABFACT, and SQA datasets across all table formats and models, including **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, **Q-2.5 7B** (Qwen-VL variants), and **GPT-5**.

Text-only VLM Pipeline: In this configuration, we disable the vision encoder and feed only the structural table format text to the VLM (Similar to handling LLM). As shown in Table 11, accuracy drops sharply compared to both multimodal VLMs and standalone LLMs. The ranking remains consistent, QWEN2.5-VL 7B > QWEN2.5-VL 3B > SMOLVLM-1.7B. HTML performs best due to its explicit tag boundaries, while LaTeX and Markdown yield low accuracy and near-zero scores on cell access. Without visual grounding, these models struggle to parse complex markup or long dependencies. Their language modules, designed for multimodal fusion, are less effective in text-only settings. TABFACT and SQA remain relatively stable thanks to their simple layouts, whereas HYBRIDQA and TOTTO are the most challenging.

LLM Pipeline: In the LLM setting, structural table text in each format is given directly to the base LLMs. Table 12 shows a similar performance hierarchy: GPT-5 > QWEN2.5-VL 7B > QWEN2.5-VL 3B > SMOLVLM-1.7B. Larger models perform better overall, though trends vary by task. Interestingly, GPT-5 underperforms on **size detection**, often overestimating column counts by treating index columns as data fields. In contrast, QWEN2.5-VL 7B shows more stable size estimation, likely due to explicit pretraining on structured text. This suggests that format sensitivity, not scale, drives accuracy on layout-oriented tasks.

Performance declines on LaTeX and Markdown, particularly for column and row retrieval, where spatial reasoning is implicit. Among datasets, TABFACT and SQA perform best, while FEVEROUS remains challenging for retrieval-heavy subtasks. TOTTO and HYBRIDQA show the widest cross-format gaps, with HTML consistently outperforming LaTeX and Markdown.

Comparative Analysis—VLM vs. LLM vs. Text-only VLM: Across all pipelines, consistent trends emerge. VLMs achieve the best overall performance, leveraging visual cues for structure-based tasks like column counting but remaining weaker on row retrieval, which depends more on semantics. LLMs excel at text-based reasoning, especially on HTML inputs, though they often overcount columns by conflating indices with data. The QWEN2.5-VL 7B model mitigates this issue, suggesting that structural token sensitivity outweighs scale. Text-only VLMs perform the weakest, confirming reliance on multimodal integration.

On HYBRIDQA, VLMs recover structure more reliably from images, while LLMs estimate rows accurately but struggle with column detection. Overall, VLMs show the most stable performance across formats by exploiting spatial cues, LLMs perform best on syntactically rich formats like HTML, and text-only VLMs add limited value without vision.

5.2 QA Results

QA results exhibit a clear difficulty gradient: binary verification tasks (FEVEROUS, TABFACT) are easiest, extractive QA (HYBRIDQA, SQA) is more demanding, and generative QA (TOTTO) is most challenging. Across all pipelines, GPT-5 achieves the highest overall performance, though the source

Format		VLM				VLM+Text			LLM			
		S-VLM	Q-2.5 3B	Q-2.5 7B	GPT-5	S-VLM	Q-2.5 3B	Q-2.5 7B	S-VLM	Q-2.5 3B	Q-2.5 7B	GPT-5
HTML	BLEU-1	3.4	**12.0**	**9.3**	8.8	2.6	10.8	7.3	6.6	5.6	8.1	**9.8**
	BLEU-2	1.9	**6.0**	**4.4**	3.8	1.5	4.5	3.2	3.0	2.9	3.7	4.6
	BLEU-3	1.2	**3.6**	**2.6**	2.2	0.9	2.6	2.0	1.7	1.8	2.2	2.7
	BLEU-4	0.8	2.5	**1.8**	1.5	0.6	1.7	1.4	1.1	1.2	1.6	1.8
Markdown	BLEU-1	**3.7**	10.5	8.6	9.7	**3.5**	**11.2**	**9.1**	**9.9**	**6.9**	**10.0**	**10.3**
	BLEU-2	**2.0**	5.4	4.2	4.3	**2.0**	**5.2**	**4.4**	**5.1**	**3.6**	**4.7**	**5.0**
	BLEU-3	**1.3**	3.3	2.5	2.5	**1.3**	**3.0**	**2.7**	**3.1**	**2.1**	**2.9**	**3.0**
	BLEU-4	**0.9**	2.3	1.7	1.4	**0.9**	**2.1**	**1.8**	**2.2**	**1.5**	**1.9**	**2.1**
LaTeX	BLEU-1	3.3	9.5	9.1	**9.0**	3.3	10.3	6.7	7.5	6.4	8.0	9.1
	BLEU-2	1.9	4.9	4.2	**4.0**	1.8	4.6	3.1	3.8	3.4	3.8	4.4
	BLEU-3	1.2	2.9	2.2	**2.3**	1.0	2.7	1.9	2.3	2.0	2.3	2.6
	BLEU-4	0.8	2.0	1.6	**1.5**	0.7	1.9	1.3	1.6	1.4	1.6	1.8

Table 4: BLUE scores for QA tasks on TOTTO datasets across all table formats and models, including **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, **Q-2.5 7B** (Qwen-VL variants), and **GPT-5**.

of its advantage varies by task and modality. Table 3 reports detailed results for FEVEROUS, HYBRIDQA, TABFACT, and SQA, while Table 4 summarizes outcomes for TOTTO.

Across models, VLM configurations, both open-weight and proprietary, perform best on non-binary QA, surpassing their LLM counterparts on extractive and compositional questions that require identifying information within specific table regions. Visual grounding enables better alignment between questions and relevant cells or columns. For binary QA, results are mixed: GPT-5 leads due to stronger text reasoning, but open VLMs often match or exceed their base LLMs, suggesting that visual cues improve factual verification. HTML and Markdown yield stable accuracy, while LaTeX remains slightly harder due to denser structure.

When visual encoders are disabled, accuracy drops sharply across datasets (Table 3). Without spatial grounding, models struggle to locate relevant regions or follow multi-step reasoning chains. Mid-sized text-only VLMs retain limited structural awareness from multimodal pretraining, but gains are small. LLMs perform reliably on binary QA, where reasoning is mainly semantic, yet lag behind VLMs on extractive and generative tasks that rely on spatial alignment. On TOTTO (Table 4), VLMs and LLMs perform comparably, indicating that while layout supports content selection, linguistic reasoning remains central to fluent generation.

Overall, these results highlight that visual grounding strengthens spatial and compositional reasoning, while textual models remain more consistent on binary and semantics-driven tasks. Together, both modalities offer complementary strengths for robust table QA.

5.3 Structure Reconstruction

We assess SR task using `GriTS`, which measures topology (TP/TR) and content (CP/CR) fidelity. GriTS aligns predicted and ground-truth grids using soft dynamic programming, while positional metrics, Cell, Row, and Column Accuracy, capture exact placements. Together, they evaluate both global layout and fine-grained reconstruction.

Intra-Format: Table 5 summarizes intra-format reconstruction results for HYBRIDQA, while Table 13 in Appendix C extends these results across all datasets and models. HTML and Markdown yield the strongest reconstructions, with HTML slightly higher due to explicit tagging. LaTeX remains the most difficult for all models because of complex syntax and layouts. Among datasets, TOTTO is hardest overall, though HTML retains stable performance.

Across models, GPT-5 achieves the highest `GriTS` and positional scores, especially on Markdown and LaTeX, producing consistent layouts with fewer missing cells. QWEN2.5-VL 7B performs reliably on LaTeX, maintaining structure even with lower cell accuracy. Markdown and LaTeX show wider gaps between topology and content scores, suggesting layout recovery exceeds text alignment, while HTML reconstructions remain the most balanced.

Format	S-VLM					Q-2.5 3B					Q-2.5 7B				
	Top	Con	Cell	Row	Col	Top	Con	Cell	Row	Col	Top	Con	Cell	Row	Col
HTML	**64.7**	**54.0**	12.3	0.4	0.1	**99.3**	**92.0**	**79.4**	**43.0**	**56.6**	**99.3**	**96.4**	**86.2**	**65.8**	**72.0**
Markdown	55.0	46.3	**20.2**	**7.3**	0.1	90.7	83.6	38.4	23.6	0.2	93.8	88.6	49.4	31.6	6.6
LaTeX	24.5	18.9	14.9	3.4	**1.8**	87.5	81.0	31.5	15.3	**1.8**	89.9	82.5	45.6	17.4	4.1

Table 5: Intra-format structure reconstruction results for HYBRIDQA, evaluated with **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, and **Q-2.5 7B** (Qwen2.5-VL variants). Models reconstruct tables from images of the same format (e.g., HTML image → HTML table). Metrics include `GriTS` topology/content scores and positional accuracy for cells, rows, and columns. Full results are in Table 13 (Appendix C).

Image	Structure	GriTS Scores		Positional Accuracy		
		Top	Con	Cell	Row	Col
HTML	HTML	**98.8**	**96.4**	**82.3**	**72.9**	**70.5**
	Md.	97.6	93.3	76.2	62.5	65.4
	LaTeX	76.1	67.8	68.5	40.3	47.4
Md.	HTML	**98.0**	93.3	81.2	66.3	63.8
	Md.	97.2	**94.9**	**83.4**	**75.2**	**74.7**
	LaTeX	61.4	53.5	40.5	24.6	23.1
LaTeX	HTML	92.4	**86.6**	**66.2**	**52.4**	**44.8**
	Md.	**93.4**	85.6	54.8	41.2	35.3
	LaTeX	88.1	78.4	61.0	42.3	24.3

Table 6: Intra- and inter-format structure reconstruction results for GPT-5. The model translates table images across formats (HTML image → Markdown/LaTeX/HTML). Metrics include `GriTS` topology/content scores and positional accuracy for cells, rows, and columns. Results are averaged over 100 samples; full results are in Table 14, Appendix C.

Inter-Format: Table 6 presents inter-format reconstruction results using GPT-5, where input and output formats differ. Accuracy decreases compared to intra-format reconstruction, but trends remain consistent. HTML is the most robust output format, maintaining high accuracy regardless of input. Markdown performs moderately well as output but inconsistently as input, while LaTeX continues to pose the greatest challenge due to strict syntax.

Occasional deviations from scaling trends appear in 14; for example, QWEN2.5-VL 3B surpasses larger models on Markdown→LaTeX `GriTS` despite lower positional accuracy, suggesting smaller models decode structure more flexibly. Cross-format success depends on both the visual clarity of the input and the syntactic constraints of the output. GPT-5 remains most stable overall, underscoring that reconstruction accuracy is influenced not only by model size but also by the trade-off between structure fidelity and decoding flexibility.

6 Conclusion

We introduced TABVERSE, a benchmark aligning identical tables across three textual formats (HTML, Markdown, LaTeX) and their rendered images. It enables direct comparison of large language and VLMs on SUC, QA, and SR, isolating the impact of representation from content. Across tasks, visual grounding enhances structure-sensitive reasoning, with VLMs excelling on non-binary QA and SUC subtasks that depend on spatial cues. Text-based reasoning remains effective for binary QA, where GPT-5 leads, while HTML, proves most stable for reconstruction and LaTeX the most challenging.

Our analysis reveals distinct model behaviors beyond QA-only settings. In SUC, GPT-5 reliably counts rows but often overestimates columns by treating index fields as data, whereas QWEN2.5-VL 7B shows better structural consistency. In Structure Reconstruction, `GriTS` and positional metrics diverge on Markdown and LaTeX, suggesting that models can recover layout but misplace content, exposing a gap between structural alignment and semantic accuracy. These differences highlight that scaling alone does not guarantee balanced improvement across structure- and content-sensitive subtasks, emphasizing the need for format-aware training objectives.

Future work should explore representation-aware decoding and Reinforcement Learning with Verifiable Rewards (RLVR) to help smaller models learn structural consistency and formatting patterns more efficiently. This can reduce column miscounts and markup errors while improving generalization across diverse table layouts. TABVERSE provides a unified foundation for studying how format and modality shape table understanding and guiding the development of more robust, structure-aware multimodal models for real-world document reasoning.

Limitations

TABVERSE is limited by its datasets, rendering setup, and evaluation scope. It integrates five English-language table benchmarks, FEVEROUS, HybridQA, SQA, TabFact, and ToTTo, which provide diverse structural and semantic coverage but restrict linguistic and cultural generalization. All tables and queries are in English, so results may not extend to other languages, scripts, or typographic conventions such as right-to-left layouts or regional numeral systems.

The visual renderings used for evaluation are generated from clean, ground-truth markup rather than real-world table images. This ensures alignment across HTML, Markdown, and LaTeX formats but omits challenges like scanning noise, handwriting, or OCR artifacts. Likewise, multi-page tables, embedded figures, and document-level variations are not simulated, meaning models are tested under idealized rather than realistic conditions.

Evaluation metrics and model coverage also impose constraints. Exact-match, BLEU, and GriTS scores capture limited semantic fidelity, and positional accuracies can over-penalize benign layout variations. Our experiments focus on a few widely used model families (SmolVLM, Qwen2.5-VL, and GPT-5), broader model diversity or document-grounded training could change observed trends. Future work will incorporate multilingual tables, noisy visual data, and document-level reasoning to improve ecological validity and generalization.

Ethical Statement and Broad Impact

TABVERSE is constructed entirely from publicly available datasets: FEVEROUS, HybridQA, SQA, TabFact, and ToTTo. All source material consists of tables and text released under open research licenses or fair-use conditions, ensuring that no proprietary or personally identifiable information is included. The benchmark is intended solely for academic research on multimodal and cross-format table understanding. While care was taken to preserve data fidelity across HTML, Markdown, and LaTeX formats, the benchmark may inherit domain or linguistic biases present in the original datasets. TABVERSE aims to advance transparent, reproducible evaluation of large language and vision–language models, and is not designed for commercial deployment or decision-making in sensitive domains.

References

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. SmolLM2: When Smol goes big – data-centric training of a small language model. *Preprint*, arXiv:2502.02737.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact Extraction and VERification Over Unstructured and Structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Alberto Barnabò. 2023. Large language models for fact-checking over tables.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020a. TabFact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, ICLR'20.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020b. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics*, EMNLP'20, pages 1026–1036, Online. Association for Computational Linguistics.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for Sequential Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. TableVQA-Bench: A visual question answering benchmark on multiple table domains. *ArXiv preprint*, abs/2404.19205.

Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. 2022. TAPEX: table pre-training via learning a neural SQL executor. In *The Tenth International Conference on Learning Representations, Virtual Event, April 25-29, 2022*, ICLR'22. OpenReview.net.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Vaibhav Srivastav, Joshua Lochner, Hugo Larcher, Mathieu Morlon, Lewis Tunstall, Leandro von Werra, and Thomas Wolf. 2025. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

OpenAI. 2025. GPT-5 system card. *OpenAI*.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, EMNLP'20, pages 1173–1186, Online. Association for Computational Linguistics.

Josselin Somerville Roberts, Tony Lee, Chi Heem Wong, Michihiro Yasunaga, Yifan Mai, and Percy Liang. 2025. Image2Struct: Benchmarking structure extraction for vision-language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Anshul Singh, Chris Biemann, and Jan Strich. 2025. MTabVQA: Evaluating multi-tabular reasoning of language models in visual space. *ArXiv preprint*, abs/2506.11684.

Ananya Singha, José Cambronero, Sumit Gulwani, Vu Le, and Chris Parnin. 2023. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLMs. In *Proceedings of the Table Representation Learning Workshop at Neural Information Processing Systems (NeurIPS) 2023*.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2023. Grits: Grid table similarity metric for table structure recognition. In *Document Analysis and Recognition - ICDAR 2023: 17th International Conference, San José, CA, USA, August 21–26, 2023, Proceedings, Part V*, page 535–549, Berlin, Heidelberg. Springer-Verlag.

Yuan Sui, Mengyu Zhou, Mingjie Zhou, Shi Han, and Dongmei Zhang. 2024a. Table Meets LLM: Can large language models understand structured table data? A benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 645–654, New York, NY, USA. Association for Computing Machinery.

Yuan Sui, Jiaru Zou, Mengyu Zhou, Xinyi He, Lun Du, Shi Han, and Dongmei Zhang. 2024b. TAP4LLM: Table provider on sampling, augmenting, and packing semi-structured data for large language model reasoning. In *Findings of the Association for Computational Linguistics*, EMNLP'24, pages 10306–10323, Miami, Florida, USA. Association for Computational Linguistics.

Pengzuo Wu, Yuhang Yang, Guangcheng Zhu, Chao Ye, Hong Gu, Xu Lu, Ruixuan Xiao, Bowen Bao, Yijing He, Liangyu Zha, Wentao Ye, Junbo Zhao, and Haobo Wang. 2025. RealHiTBench: A comprehensive realistic hierarchical table benchmark for evaluating LLM-based table analysis. In *Findings of the Association for Computational Linguistics*, ACL'25, pages 7105–7137, Vienna, Austria. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *ArXiv preprint*, abs/2412.15115.

Han Zhang, Yuheng Ma, and Hanfang Yang. 2025. ALTER: Augmentation for large-table-based reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 179–198, Albuquerque, New Mexico. Association for Computational Linguistics.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, Houqiang Li, and Can Huang. 2024. TabPedia: Towards comprehensive visual table understanding with concept synergy. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, December 10 - 15, 2024*, NeurIPS'24, Vancouver, BC, Canada.

Junnan Zhu, Jingyi Wang, Bohan Yu, Xiaoyu Wu, Junbo Li, Lei Wang, and Nan Xu. 2025. TableEval: A real-world benchmark for complex, multilingual, and multi-structured table question answering. *ArXiv preprint*, abs/2506.03949.

A Prompts

Task	Prompt
Table Partition	What is the **first cell value** (not including headers) of the given table? What is the **last cell value** (not including headers) of the given table? Answer questions one by one and use | to split the answer. Answer the question without having any introduction or explanations.
First Cell	What is the **first cell value** (not including headers) of the given table? Answer the question without having any introduction or explanations.
Last Cell	What is the **last cell value** (not including headers) of the given table? Answer the question without having any introduction or explanations.
Size Detection	How many rows in the table? How many columns in the table? Answer the questions one by one and use | to split the answer. Answer the question without having any introduction or explanations.
No. of Rows	How many rows in the table? Answer the question without having any introduction or explanations.
No. of Columns	How many columns in the table? Answer the question without having any introduction or explanations.
Cell Lookup	Row/column indices start at 0 (top-left is 0|0). What is the position of the cell value `{cell_value}`? Use row index and column index to answer. Use | to split the answer. Answer the question without having any introduction or explanations.
Reverse Lookup	Row/column indices start at 0 (top-left is 0|0). What is the cell value of row index `{reverse_lookup_row}`, column index `{reverse_lookup_col}`? Only output the cell value without other information. Answer the question without having any introduction or explanations.
Column Retrieval	Row/column indices start at 0 (top-left is 0|0). What is the column name with the index `{column_idx}` of the given table image? Only give the column name without any explanation. Answer the question without having any introduction or explanations.
Row Retrieval	Row/column indices start at 0 (top-left is 0|0). What are the cell values of the `{row_idx}` row in following table? Only list the cell values one by one using | to split the answers. Answer the question without having any introduction or explanations.

Table 7: SUC task prompts used across modalities and formats.

QA Type	Prompt
General QA	Look at the given image and answer the following question directly. Do not include introductions, explanations, or extra text. Provide only the exact, precise final answer. `{query}`
Binary QA	Look at the given image and answer the following question with only a single digit: 1 if the statement is true, 0 if the statement is false. Do not include any explanations or extra text. `{query}`

Table 8: QA task prompts reused across modalities and formats.

Format	Prompt
HTML	Generate the complete HTMLcode that exactly represents this image. Provide only the code without any explanations.
LaTeX	Generate the complete LaTeX code that exactly represents this image. Provide only the code without any explanations.
Markdown	Generate the complete Markdown code that exactly represents this image. Provide only the code without any explanations.

Table 9: Structure Reconstruction task prompts used across different table formats.

B Figures

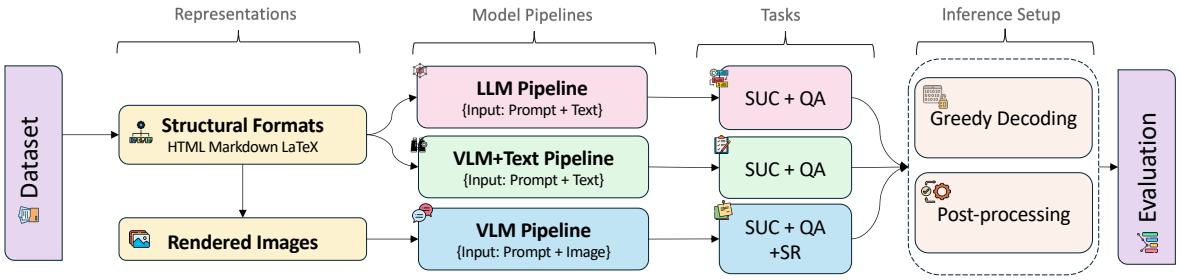

Figure 2: End-to-end workflow of TABVERSE. Each table instance is represented in three textual formats (HTML, Markdown, LaTeX) and respective rendered images. These are processed through three model pipelines: **LLM**, **VLM+Text**, and **VLM**, covering the **SUC**, **QA**, and **GEN** tasks. Inference uses greedy decoding followed by post-processing, with results aggregated for evaluation.

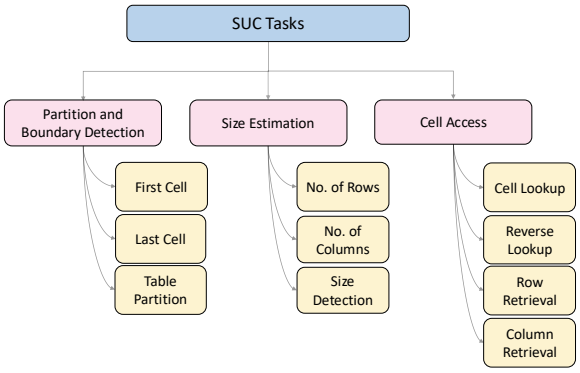

Figure 3: Taxonomy of Structural Understanding Capability (SUC) tasks. Ten tasks are grouped into partitioning, size estimation, value lookup, and row/column retrieval.

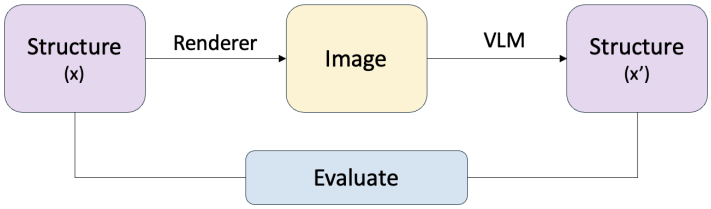

Figure 4: Pipeline for structure reconstruction. A ground-truth table x is rendered into an image, the model predicts a structure x', and evaluation compares x' with x.

C Result Tables

765

Models	Dataset	Formats	Detection			Size Estimation			Cell Access			
			T.P.	F.C.	L.C.	S.D.	# Rows	# Cols	C.L.	R.L.	C.R.	R.R.
S-VLM	Feverous	Html	0.0	17.3	34.3	1.7	21.3	35.7	4.9	21.6	33.2	0.0
		Latex	0.0	16.3	32.5	0.3	23.5	24.2	6.3	18.4	32.2	0.0
		Markdown	0.5	27.1	40.3	0.9	23.3	27.7	3.8	19.7	35.3	0.0
	TabFact	Html	0.1	23.5	32.3	1.1	28.8	30.5	1.9	14.1	28.5	0.0
		Latex	0.0	9.8	33.6	0.7	33.7	27.4	2.7	11.7	21.5	0.0
		Markdown	0.4	27.0	48.0	0.9	34.1	22.7	2.6	12.6	26.9	0.0
	SQA	Html	0.0	34.5	37.8	0.7	24.3	23.0	1.4	16.2	28.4	0.0
		Latex	0.0	20.9	38.5	0.0	25.0	12.2	2.7	12.8	25.0	0.0
		Markdown	0.7	34.5	43.9	0.0	27.7	21.6	0.7	17.6	29.7	0.0
	ToTTo	Html	0.1	15.4	18.1	1.6	20.0	28.5	5.4	13.5	25.7	0.1
		Latex	0.0	15.3	18.4	0.1	14.8	20.5	3.7	10.4	20.2	0.0
		Markdown	0.0	20.8	25.3	0.7	17.4	23.4	4.0	14.6	21.4	0.0
Q-2.5 3B	Feverous	Html	3.3	57.0	49.8	32.3	77.3	24.4	9.8	31.2	35.9	9.1
		Latex	6.0	52.6	45.5	43.3	71.3	62.0	10.6	33.1	62.8	4.9
		Markdown	9.9	52.0	49.4	49.8	69.8	65.2	8.7	34.0	60.2	7.7
	TabFact	Html	2.6	78.2	65.5	49.3	82.7	49.3	11.4	33.1	38.8	17.0
		Latex	9.2	76.2	62.1	48.5	76.5	55.8	15.8	39.3	63.2	10.7
		Markdown	15.4	61.5	72.4	54.1	67.0	73.0	13.5	39.1	67.0	20.9
	SQA	Html	3.4	64.2	60.8	41.9	68.2	40.5	6.8	27.0	35.1	14.9
		Latex	7.4	62.2	58.8	36.5	66.2	53.4	12.8	33.1	56.8	8.1
		Markdown	13.5	58.8	65.5	37.2	65.5	68.2	8.8	37.2	61.5	11.5
	ToTTo	Html	1.7	43.1	35.3	27.1	55.9	36.3	9.4	22.7	32.4	8.0
		Latex	5.1	47.4	31.8	26.2	45.4	57.5	8.8	25.5	50.2	6.1
		Markdown	7.8	29.5	39.4	28.8	42.5	59.2	5.8	21.4	48.8	7.8
Q-2.5 7B	Feverous	Html	11.52	55.6	51.5	22.0	97.2	30.9	19.5	27.9	34.3	13.5
		Latex	13.5	50.2	47.7	75.0	91.6	76.4	23.0	33.9	64.0	10.6
		Markdown	17.6	57.8	52.6	59.4	94.5	69.1	22.4	28.7	49.6	14.3
	TabFact	Html	34.3	84.7	74.3	30.3	97.8	53.6	20.8	26.4	34.9	26.4
		Latex	24.7	74.9	65.8	81.4	95.2	84.4	24.9	28.5	65.9	21.7
		Markdown	25.1	69.9	71.9	66.6	95.7	78.2	22.3	21.8	50.1	27.3
	SQA	Html	18.9	66.2	58.8	23.6	97.3	47.3	14.2	25.7	29.1	13.5
		Latex	20.9	60.8	59.5	77.0	89.2	75.7	18.2	27.7	62.8	14.9
		Markdown	19.6	63.5	64.2	67.6	91.9	76.4	18.9	24.3	53.4	15.5
	ToTTo	Html	8.5	45.0	36.9	27.5	87.8	45.3	16.2	19.9	28.2	13.0
		Latex	9.2	43.7	34.3	53.9	75.6	68.7	16.3	20.3	52.2	10.1
		Markdown	9.4	42.8	40.4	37.2	84.9	58.2	16.5	16.1	33.3	13.1
GPT-5	Feverous	Html	29.9	35.6	69.1	33.1	96.1	52.6	32.9	35.9	57.8	51.2
		Latex	34.9	67.7	66.1	77.0	94.5	82.4	55.0	51.7	86.4	51.5
		Markdown	38.1	62.2	69.1	59.4	94.0	76.2	47.1	51.7	82.4	57.2
	TabFact	Html	67.2	59.6	93.4	55.0	97.8	61.2	38.1	50.1	64.9	80.9
		Latex	40.8	55.2	89.1	76.0	97.4	76.8	53.4	64.3	87.8	69.8
		Markdown	50.5	63.6	86.5	72.0	97.8	77.0	41.4	56.8	83.6	74.6
	SQA	Html	60.1	56.1	92.6	43.2	99.3	56.1	31.8	53.4	61.5	75.0
		Latex	45.9	61.5	91.9	66.9	95.3	82.4	45.3	59.5	82.4	62.8
		Markdown	51.4	58.1	91.2	54.7	93.9	68.2	37.2	58.1	81.8	52.0
	ToTTo	Html	30.5	37.2	60.6	37.2	85.4	55.0	33.7	35.8	64.4	46.5
		Latex	26.4	49.9	55.2	54.5	74.2	71.9	38.4	41.6	79.7	36.3
		Markdown	27.0	42.2	59.6	40.7	77.5	62.6	29.5	40.2	73.6	40.7

Table 10: Performance comparison of VLMs on the SUC benchmark across three table formats (HTML, LaTeX, and Markdown) and datasets. Models include **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, and **Q-2.5 7B** (both Qwen-VL variants), and **GPT-5**. Metrics span three subtasks: *Partition and Boundary Detection* (T.P.=Table Partition, F.C.=First Cell, L.C.=Last Cell), *Size Estimation* (S.D.=Size Detection, No. of Rows, No. of Cols), and *Cell Access* (C.L.=Cell Lookup, R.L.=Reverse Lookup, C.R.=Column Retrieval, R.R.=Row Retrieval). See Table 2 for HYBRIDQA Results.

Models	Dataset	Formats	Detection			Size Estimation			Cell Access			
			T.P.	F.C.	L.C.	S.D.	# Rows	# Cols	C.L.	R.L.	C.R.	R.R.
S-VLM	Feverous	Html	0.0	28.5	5.4	0.0	19.5	6.8	2.0	4.3	14.2	0.2
		Latex	0.0	13.4	6.3	0.2	16.7	7.1	2.4	2.7	7.6	0.0
		Markdown	0.0	19.2	4.9	0.0	16.7	7.9	1.9	2.7	15.3	0.2
	TabFact	Html	0.0	35.7	7.7	0.0	13.9	8.1	1.9	3.1	10.8	0.0
		Latex	0.0	11.9	8.2	0.1	9.6	9.7	1.5	3.0	8.9	0.0
		Markdown	0.0	10.7	3.2	0.0	9.9	9.5	2.8	1.6	9.9	0.0
	SQA	Html	0.0	38.5	7.4	0.0	19.6	6.1	2.0	5.4	12.8	0.0
		Latex	0.0	15.1	6.8	0.0	11.6	4.8	2.7	2.7	9.6	0.0
		Markdown	0.0	17.6	9.5	0.0	14.2	4.7	2.7	2.0	16.2	0.0
	HybridQA	Html	0.0	32.6	4.6	0.0	15.2	0.0	2.0	3.2	7.7	0.0
		Latex	0.0	11.1	3.1	0.0	10.9	0.3	1.0	2.4	1.1	0.0
		Markdown	0.0	11.8	1.6	0.0	12.0	0.3	3.3	1.1	3.6	0.0
	ToTTo	Html	0.0	28.3	3.4	0.0	16.2	13.9	2.2	3.5	11.3	0.3
		Latex	0.0	16.3	3.1	0.1	10.8	10.2	1.7	3.6	5.9	0.1
		Markdown	0.0	15.3	2.3	0.0	14.2	7.0	2.2	1.7	9.8	0.1
Q-2.5 3B	Feverous	Html	4.4	41.9	16.4	2.7	78.7	12.0	11.5	25.2	33.5	2.4
		Latex	2.7	17.6	11.2	0.3	47.1	1.1	7.2	18.7	40.0	2.5
		Markdown	5.0	28.8	16.7	8.8	69.3	14.6	9.0	17.2	26.6	0.6
	TabFact	Html	3.9	25.5	8.6	9.3	76.2	18.8	9.5	16.1	28.5	1.8
		Latex	2.6	4.1	3.9	0.9	43.1	1.4	5.9	11.8	29.3	0.7
		Markdown	6.2	11.8	13.9	9.1	61.6	21.4	7.3	13.1	23.2	3.4
	SQA	Html	4.1	31.1	7.4	1.4	70.3	11.5	6.8	21.6	27.7	2.7
		Latex	4.7	15.5	8.1	0.7	59.5	2.0	6.8	15.5	27.0	2.0
		Markdown	4.1	25.7	10.1	6.1	68.2	10.1	10.1	14.2	20.3	2.7
	HybridQA	Html	3.0	41.5	8.1	0.8	66.0	4.0	7.9	19.3	38.9	5.5
		Latex	3.0	17.7	2.2	0.5	52.7	2.2	4.4	14.1	40.0	2.5
		Markdown	2.8	33.2	5.1	2.2	31.4	10.1	5.2	12.4	30.9	3.1
	ToTTo	Html	7.0	31.4	11.4	10.7	60.1	25.0	15.8	21.4	26.1	4.9
		Latex	4.1	8.4	5.9	0.9	39.0	5.4	8.1	14.3	28.3	2.2
		Markdown	5.1	16.9	9.1	10.9	47.9	25.8	10.5	15.2	21.7	3.5
Q-2.5 7B	Feverous	Html	15.3	75.7	37.2	50.7	69.0	86.8	21.1	33.1	56.9	15.0
		Latex	3.8	3.0	36.5	0.2	23.5	0.2	10.4	24.7	29.6	8.8
		Markdown	7.4	41.9	37.6	17.6	72.9	27.6	11.7	23.6	30.6	9.8
	TabFact	Html	15.7	68.7	26.6	42.5	68.6	78.3	20.6	26.8	47.1	18.5
		Latex	2.0	1.4	49.1	0.0	26.5	0.0	8.5	21.1	25.9	21.8
		Markdown	12.7	27.4	45.1	9.9	56.8	15.0	9.5	20.4	23.1	27.8
	SQA	Html	12.8	87.2	38.5	66.9	87.8	98.0	17.6	24.3	52.0	29.7
		Latex	3.4	6.1	40.5	0.0	27.0	0.0	7.4	17.6	29.7	20.3
		Markdown	9.5	60.8	38.5	6.1	92.6	14.9	9.5	20.3	32.4	16.2
	HybridQA	Html	10.4	87.5	22.3	41.9	73.3	93.7	16.4	29.5	59.6	32.9
		Latex	3.1	0.8	15.3	0.3	31.2	1.2	6.5	21.3	32.3	21.0
		Markdown	6.3	21.8	12.6	7.5	54.3	17.1	6.1	21.1	28.7	24.0
	ToTTo	Html	15.7	68.7	26.6	42.5	68.6	78.3	20.6	26.8	47.1	18.5
		Latex	2.8	1.9	27.7	1.3	29.9	1.1	10.7	20.7	24.7	9.6
		Markdown	9.2	41.1	20.3	13.7	52.1	16.9	12.5	21.9	23.7	11.9

Table 11: Performance comparison of VLMs+Text on the SUC benchmark across three table formats (HTML, LaTeX, and Markdown) and datasets. Models include **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, and **Q-2.5 7B** (both Qwen-VL variants). Metrics span three subtasks: *Partition and Boundary Detection* (T.P.=Table Partition, F.C.=First Cell, L.C.=Last Cell), *Size Estimation* (S.D.=Size Detection, No. of Rows, No. of Cols), and *Cell Access* (C.L.=Cell Lookup, R.L.=Reverse Lookup, C.R.=Column Retrieval, R.R.=Row Retrieval).

Models	Dataset	Formats	Detection			Size Estimation			Cell Access			
			T.P.	F.C.	L.C.	S.D.	# Rows	# Cols	C.L.	R.L.	C.R.	R.R.
S-LM	Feverous	Html	0.0	26.3	23.8	2.5	20.9	8.2	3.1	3.5	6.8	0.0
		Latex	0.0	11.5	13.4	2.8	22.7	7.9	3.5	2.2	7.2	0.0
		Markdown	0.0	12.8	10.1	0.0	10.1	9.0	3.5	3.3	6.5	0.0
	TabFact	Html	0.4	29.2	23.5	0.3	13.0	10.4	1.1	2.2	3.6	0.0
		Latex	0.0	15.0	20.4	1.1	19.1	10.5	0.9	2.3	4.2	0.0
		Markdown	0.0	9.2	14.2	0.0	8.0	6.6	1.1	1.2	2.4	0.0
	SQA	Html	0.7	38.9	19.4	0.0	13.2	6.9	2.8	3.5	4.2	0.0
		Latex	0.0	23.6	18.9	0.7	18.9	6.1	1.4	4.1	8.1	0.0
		Markdown	0.0	12.2	16.9	0.0	10.8	5.4	1.4	4.7	6.1	0.0
	HybridQA	Html	0.2	42.7	13.9	0.1	11.0	0.0	2.9	2.5	1.9	0.0
		Latex	0.0	18.2	5.6	0.2	13.1	0.1	1.6	1.4	0.3	0.0
		Markdown	0.0	18.9	6.0	0.0	11.5	0.1	0.9	1.0	0.3	0.0
	ToTTo	Html	0.0	27.3	14.3	4.7	14.1	17.5	3.3	2.8	12.9	0.0
		Latex	0.0	15.4	11.7	2.4	11.6	10.7	1.7	1.7	6.4	0.0
		Markdown	0.0	14.8	9.2	0.0	8.4	8.6	2.7	2.1	4.8	0.0
Q-2.5 3B	Feverous	Html	5.4	26.9	22.4	13.9	78.6	53.1	19.8	23.3	44.7	1.3
		Latex	1.4	32.4	35.6	1.7	63.8	8.8	6.0	18.3	39.7	1.4
		Markdown	1.6	15.4	26.3	9.3	71.0	18.6	11.7	16.1	34.0	0.3
	TabFact	Html	4.2	31.8	27.8	13.8	75.4	50.9	16.1	15.5	32.6	1.1
		Latex	0.9	14.9	43.8	4.5	63.6	35.4	7.8	12.8	34.5	0.9
		Markdown	1.1	22.4	27.2	4.7	72.0	8.6	10.0	11.4	27.8	0.3
	SQA	Html	2.0	29.1	24.3	10.1	75.0	50.0	18.9	14.9	35.1	0.7
		Latex	2.7	35.1	43.9	2.7	70.3	27.0	8.1	18.2	33.1	0.7
		Markdown	1.4	21.6	25.7	2.0	75.7	10.1	12.8	16.2	32.4	0.7
	HybridQA	Html	2.4	45.3	27.1	5.6	61.9	70.5	17.9	19.9	45.7	1.4
		Latex	0.3	33.5	32.2	1.6	69.9	14.8	6.0	13.4	41.5	0.8
		Markdown	0.4	36.4	27.2	8.4	61.8	17.9	11.4	12.9	31.4	0.4
	ToTTo	Html	4.8	26.5	19.8	14.4	63.7	41.5	20.9	18.8	28.1	4.3
		Latex	1.4	24.7	27.7	1.8	46.4	10.5	8.0	13.6	38.2	2.6
		Markdown	0.8	20.7	19.3	9.2	50.2	22.3	12.4	11.6	29.0	1.3
Q-2.5 7B	Feverous	Html	13.2	89.6	34.5	63.3	95.6	78.7	34.0	31.0	48.5	17.6
		Latex	8.2	8.0	45.5	0.8	76.4	0.3	15.4	15.0	33.5	12.0
		Markdown	4.3	51.3	42.7	15.1	94.2	45.5	20.6	21.6	33.1	9.3
	TabFact	Html	17.4	98.9	47.7	81.4	97.2	94.2	26.1	21.9	46.9	30.8
		Latex	7.4	0.8	58.8	0.4	77.4	0.0	11.6	12.3	24.3	19.1
		Markdown	4.6	51.1	48.6	9.6	87.2	24.3	16.4	16.9	19.5	12.8
	SQA	Html	18.2	95.9	43.2	65.5	93.2	92.6	29.1	23.6	38.5	33.1
		Latex	18.2	10.1	54.1	0.0	94.6	0.0	6.1	17.6	28.4	19.6
		Markdown	5.4	54.1	43.9	8.8	93.2	13.5	16.2	18.2	20.9	16.2
	HybridQA	Html	10.4	97.4	26.4	57.6	93.3	88.0	37.4	25.3	48.7	35.4
		Latex	0.9	9.2	31.4	0.8	76.9	0.3	13.4	11.9	30.2	12.8
		Markdown	1.7	63.4	19.1	8.8	71.9	39.4	19.6	17.8	26.8	12.5
	ToTTo	Html	12.9	87.7	27.1	57.6	90.8	82.1	29.3	28.7	51.4	21.6
		Latex	6.4	6.8	36.8	1.2	63.4	0.9	13.8	17.0	30.3	13.2
		Markdown	1.8	57.0	24.6	14.6	76.9	53.9	17.6	19.2	26.8	11.6
GPT-5	Feverous	Html	80.8	95.7	87.1	25.2	99.7	3.8	16.1	47.1	66.3	95.7
		Latex	7.1	2.2	82.5	0.0	47.1	0.0	3.8	25.5	20.8	39.2
		Markdown	60.2	63.8	80.2	0.8	97.2	0.8	16.7	34.5	35.9	67.6
	TabFact	Html	99.5	100.0	99.9	51.5	99.9	27.0	18.8	48.8	76.2	98.8
		Latex	11.4	2.0	99.2	0.0	77.3	0.0	2.8	18.6	14.5	55.1
		Markdown	70.9	59.1	99.1	1.6	98.2	1.6	4.9	25.9	21.9	74.1
	SQA	Html	100.0	100.0	100.0	17.6	99.3	5.4	6.8	39.9	64.2	99.3
		Latex	10.8	4.1	98.0	0.0	66.2	0.0	3.4	15.5	15.5	48.6
		Markdown	77.7	64.2	98.0	0.7	98.0	0.0	6.8	29.7	23.6	72.3
	HybridQA	Html	93.5	98.2	95.9	19.0	100.0	3.9	7.9	45.9	75.9	98.9
		Latex	15.0	5.2	91.8	0.0	82.5	0.0	2.0	24.6	22.6	53.7
		Markdown	64.9	58.2	91.1	0.1	99.5	0.1	11.6	32.1	30.8	79.6
	ToTTo	Html	76.3	94.5	86.1	19.3	98.8	4.2	20.2	42.4	69.6	96.2
		Latex	8.0	4.0	80.1	0.0	47.3	0.0	5.5	20.2	25.7	47.0
		Markdown	54.1	67.3	78.1	2.4	88.7	8.9	16.4	28.7	35.8	70.1

Table 12: Performance comparison of LLMs on the SUC benchmark across three table formats (HTML, LaTeX, and Markdown) and datasets. Models include **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, and **Q-2.5 7B** (both Qwen-VL variants), and **GPT-5**. Metrics span three subtasks: *Partition and Boundary Detection* (T.P.=Table Partition, F.C.=First Cell, L.C.=Last Cell), *Size Estimation* (S.D.=Size Detection, No. of Rows, No. of Cols), and *Cell Access* (C.L.=Cell Lookup, R.L.=Reverse Lookup, C.R.=Column Retrieval, R.R.=Row Retrieval).

Dataset	Format	S-VLM									Q-2.5 3B									Q-2.5 7B								
		GriTS Scores						Pos Acc			GriTS Scores						Pos Acc			GriTS Scores						Pos Acc		
		Top	Con	TP	TR	CP	CR	Cell	Row	Col	Top	Con	TP	TR	CP	CR	Cell	Row	Col	Top	Con	TP	TR	CP	CR	Cell	Row	Col
Feverous	HTML	64.4	51.0	74.4	58.4	58.5	46.2	17.8	1.5	0.2	98.3	90.0	98.7	98.1	90.4	89.8	81.1	49.2	63.9	98.3	92.1	99.2	97.9	92.6	91.8	85.0	66.9	69.1
	Md	72.4	59.5	80.3	67.9	65.5	55.6	38.0	16.9	1.8	86.5	76.7	98.2	79.3	85.7	70.2	23.9	10.0	0.1	89.2	82.9	97.3	83.2	90.4	77.2	33.9	23.1	13.1
	LaTeX	22.0	16.1	20.5	24.8	15.1	18.1	15.0	4.1	1.8	89.6	77.8	95.2	85.6	82.6	74.3	53.3	32.7	3.1	87.3	73.2	93.1	83.1	78.1	69.7	51.4	24.3	7.3
H-QA	HTML	64.7	54.0	74.1	58.8	61.4	48.9	12.3	0.4	0.1	99.3	92.0	99.3	99.3	92.1	92.1	79.4	43.0	56.6	99.3	96.4	99.7	99.0	96.8	96.1	86.2	65.8	72.0
	Md	55.0	46.3	61.2	51.3	50.7	43.0	20.2	7.3	0.1	90.7	83.6	98.3	86.3	89.2	79.5	38.4	23.6	0.2	93.8	88.6	99.2	89.6	93.7	84.5	49.4	31.6	6.6
	LaTeX	24.5	18.9	24.5	26.5	18.4	20.1	14.9	3.4	1.8	87.5	81.0	96.1	81.2	88.4	75.2	31.5	15.3	1.8	89.9	82.5	97.4	84.4	89.0	77.5	45.6	17.4	4.1
SQA	HTML	71.6	59.5	77.9	68.5	64.4	56.4	12.1	1.6	0.0	98.3	90.1	98.2	98.4	90.0	90.2	76.0	45.7	59.6	99.3	96.5	99.6	99.0	96.8	96.2	89.6	82.4	76.5
	Md	74.8	64.3	80.9	72.1	68.7	61.1	33.4	14.9	0.8	90.2	84.2	99.9	84.2	91.0	78.6	20.3	7.0	0.4	92.4	89.4	98.6	87.1	95.5	84.2	26.1	17.5	7.9
	LaTeX	15.2	13.2	15.2	15.8	13.3	13.5	10.3	5.0	0.5	91.5	85.1	97.4	86.6	90.8	80.5	37.5	22.4	4.4	92.3	87.1	97.5	88.3	92.2	82.9	45.8	25.4	4.3
TabFact	HTML	78.8	65.8	83.6	77.4	69.5	64.3	14.8	1.2	0.2	99.4	90.6	99.4	99.4	90.6	90.6	80.9	42.6	64.5	99.1	88.4	99.3	98.9	88.6	88.3	79.4	58.2	65.0
	Md	74.9	63.5	81.1	71.6	67.7	60.7	39.4	19.3	3.9	92.7	85.0	99.1	88.2	90.1	80.8	31.7	12.2	0.1	95.3	92.6	99.5	91.7	96.8	89.1	47.6	33.6	30.8
	LaTeX	33.1	26.4	32.9	35.0	25.7	27.7	20.0	5.8	2.1	92.7	87.1	98.1	88.3	92.2	82.7	34.3	16.6	3.2	94.1	89.4	99.3	89.8	94.4	85.1	39.2	18.1	5.3
ToTTo	HTML	56.4	41.6	65.8	52.2	48.0	38.0	16.0	0.4	0.2	93.8	81.8	96.4	93.1	83.7	81.2	69.5	36.4	44.6	92.9	84.0	97.7	92.2	85.9	83.3	73.9	52.2	50.3
	Md	55.0	39.6	62.1	53.2	43.4	37.4	27.0	6.6	1.6	85.0	71.7	92.0	82.4	76.5	68.9	31.1	12.5	4.0	88.4	81.5	96.5	82.8	89.0	76.2	31.8	16.9	11.6
	LaTeX	16.7	11.8	15.7	19.1	11.0	13.3	10.7	2.1	1.2	80.4	66.5	87.3	76.9	71.8	63.1	40.4	17.7	4.3	79.9	68.5	86.5	75.5	74.0	64.6	38.3	15.1	6.3

Table 13: Intra-format structure reconstruction results across all datasets and table formats, evaluated for **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, and **Q-2.5 7B** (both Qwen-VL variants). Models reconstruct structured representations (e.g., HTMLimage → HTMLtable). **GriTS** scores include *Top/Con* (topology/content) and their precision/recall components (*TP/TR, CP/CR*). **Pos Acc** reports positional accuracy across cells, rows, and columns.

Model	Image Format	Structure Format	GriTS Scores						Pos Acc		
			Top	Con	TP	TR	CP	CR	Cell	Row	Col
S-VLM	Html	Markdown	84.8	71.2	95.0	77.5	79.4	65.1	17.2	0.0	0.0
		Latex	13.2	10.3	22.7	13.9	10.6	10.7	7.6	0.0	3.3
	Markdown	Html	53.1	43.2	59.6	48.9	48.1	39.9	17.3	3.3	0.0
		Latex	12.5	8.0	12.7	12.2	8.2	7.8	5.5	0.3	0.0
	Latex	Html	78.1	56.2	83.1	74.9	59.7	53.7	26.1	0.8	0.0
		Markdown	91.4	69.1	97.4	89.2	70.7	68.1	52.4	28.7	7.3
Q-2.5 3B	Html	Markdown	93.2	86.7	100.0	87.6	93.0	81.6	30.3	21.4	11.8
		Latex	98.4	95.2	98.8	98.1	95.6	94.9	83.3	70.9	61.7
	Markdown	Html	85.5	76.8	96.6	80.9	87.4	72.4	20.2	6.7	10.9
		Latex	90.3	78.4	95.5	88.0	83.2	76.3	29.3	8.6	18.8
	Latex	Html	80.6	66.7	85.4	77.3	70.0	64.7	54.1	32.9	1.0
		Markdown	91.2	79.0	100.0	84.5	86.9	73.0	31.3	8.5	0.0
Q-2.5 7B	Html	Markdown	77.7	75.7	80.0	75.8	77.9	73.8	50.6	45.2	40.4
		Latex	90.6	81.7	91.7	91.2	83.1	81.3	71.9	31.6	51.2
	Markdown	Html	96.5	88.3	99.8	93.7	91.5	85.7	59.0	36.6	30.4
		Latex	66.4	55.5	70.0	63.4	58.3	53.1	31.0	10.8	18.5
	Latex	Html	94.0	83.6	98.0	90.6	87.2	80.7	54.1	32.9	1.0
		Markdown	93.3	82.5	97.3	90.2	85.9	79.7	20.6	8.5	1.3
GPT-5	Html	Html	98.8	96.4	99.9	97.4	97.5	95.4	82.3	72.9	70.5
		Markdown	97.6	93.3	100.0	95.7	95.5	91.6	76.2	62.5	65.4
		Latex	76.1	67.8	76.0	76.9	67.9	67.6	68.5	40.3	47.4
	Markdown	Html	98.0	93.3	100.0	96.4	95.1	91.9	81.2	66.3	63.8
		Markdown	97.2	94.9	97.1	98.0	94.8	95.5	83.4	75.2	74.7
		Latex	61.4	53.5	64.5	59.5	56.3	51.5	40.5	24.6	23.1
	Latex	Html	92.4	86.6	97.7	88.5	89.5	84.5	66.2	52.4	44.8
		Markdown	93.4	85.6	97.3	90.4	88.5	82.9	54.8	41.2	35.3
		Latex	88.1	78.4	93.7	84.6	81.9	76.0	61.0	42.3	24.3

Table 14: Inter-format structure reconstruction results, where models translate table images across representation formats (e.g., HTMLImage → Markdown/LaTeX structured text). Evaluations include inter-format transformations for **S-VLM** (SmolVLM-1.7B), **Q-2.5 3B**, **Q-2.5 7B** (Qwen-VL variants), and both inter-format and intra-format for **GPT-5**, computed over a 100-sample subset. **GriTS** scores capture *Top/Con* (topology/content) and their precision/recall components (*TP/TR, CP/CR*), while **Pos Acc** reports positional accuracy across cells, rows, and columns.

D Error Analysis

D.1 QA Task

We perform qualitative error analysis on the *QA task* using a subset of 30 representative images sampled from the SQA, HybridQA, and ToTTo datasets to identify the main factors contributing to lower quantitative performance. Table 15 summarizes the major error categories observed across HTML, Markdown, and LaTeX formats, along with illustrative examples. Each case exposes a distinct failure mode, including structural misalignment, symbol confusion, and relational or grounding errors. The subsequent subsections analyze these cases in detail, demonstrating how format-specific serialization and table structure influence model reasoning and answer accuracy.

Error	Query	Label	Prediction	Image
Entity Fragmentation	How many craters were named after Americans?	2	3	5
Numeric Symbol Error	Which school had the fewest wins?	Penn State	Connecticut	6
Delimiter Drift	Which airport handled the most passengers?	Netherlands – Amsterdam	Netherlands – Amsterdam	7
Cross-column Linking	What position was the player whose MLS team rebranded in 2010?	F	D	8
Cell-to-Row Grounding (ToTTo)	Produce a one-sentence description for each highlighted cells ([[15, 0], [15, 1]]) of the table.	Burton's population was 1,452 at the 2010 census.	Row 16, Column 1 (Census): Est. 2017. Row 16, Column 2 (Pop.): 1,454.	9
Relational Composition (ToTTo)	Produce a one-sentence description for each highlighted cells ([[1, 3], [2, 0], [2, 1], [2, 3]]) of the table.	Britain Simons appeared in another Lifetime Television film (2015).	The Notes cell for row 1 lists "Lifetime Television.[2,0]: The Year cell for row 2 shows "2016."[2,1]: The Title cell for row 2 is "Holidays."[2,3]: The Notes cell for row 2 states "XYZ Films"	10
Row-ordering Error	Which match was played the earliest in the tournament?	Liverpool vs. West Bromwich Albion	Cardiff City vs. Southampton	11

Table 15: Representative error categories in the QA task across formats, with example questions, gold labels, and model predictions.

D.1.1 Entity Fragmentation

In Figure 5, the model overestimates the number of craters named after Americans, predicting three instead of the correct two. This reflects an *entity fragmentation* error, where HTMLmarkup splits descriptive text such as "William Faulkner, American author" into separate tokens or spans. The model then misinterprets the comma-delimited fragments ("Faulkner" and "American author") as distinct entities, inflating the nationality count. In contrast, both Markdown and LaTeXserialize the same phrase as a continuous text sequence, preserving the intended one-to-one relation between person and nationality and therefore produce the correct prediction.

D.1.2 Numeric Symbol Error

In Figure 6, the question asks which school had the fewest wins in the *Record (conference)* column, where the left-hand value represents the number of wins (e.g., 15–4). The HTMLand Markdown versions correctly identify this column and return Penn State (15–4). In contrast, the LaTeXversion introduces ambiguity due to duplicated header tokens "Record (conference)" vs. "CWS record", math-style dashes, and nested parentheses. These typographic artifacts confuse the model's column association, leading it to align "wins" with the *CWS record* field instead. Because Connecticut's entry (1–2) appears numerically

	Crater	Diameter (km)	Approval Year	Eponym	Ref
0	Faulkner	168	2012	William Faulkner , American author	WGPSN
1	Fet	79	1985	Afanasy Fet , Russian poet	WGPSN
2	Firdousi	98	2010	Hakim Ferdowsi , Persian poet	WGPSN
3	Flaubert	95	1985	Gustave Flaubert , French author	WGPSN
4	Flaiano	43	2013	Ennio Flaiano , Italian author	WGPSN
5	Fonteyn	29	2012	Margot Fonteyn , English ballet dancer	WGPSN
6	Fuller	26.97	2013	Richard Buckminster Fuller , American engineer and architect	WGPSN
7	Futabatei	57	1976	Futabatei Shimei , Japanese author	WGPSN

Figure 5: Entity fragmentation error.

Table 1:

	School	Conference	Record (conference)	Head coach	CWS appearances	CWS best finish	CWS record
0	Arizona	nan	36-8 (N/A)	Frank Sancet	4 (last: 1958)	2nd (1956)	7-8
1	Clemson	ACC	23 -6 (11-5, 0 GB)	Bill Wilhelm	1 (last: 1958)	5th (1958)	1-2
2	Colorado State	nan	24-5 (N/A)	Pete Butler	5 (last: 1958)	5th (1955)	2-10
3	Connecticut	nan	20-1 (N/A)	J. Orlean Christian	1 (last: 1957)	5th (1957)	1-2
4	Fresno State	nan	38-11 (N/A)	Pete Beiden	0 (last: none)	none	0-0
5	Oklahoma State	Big 8	22-4 (17-3, 0 GB)	Toby Greene	2 (last: 1955)	3rd (1955)	5–4
6	Penn State	nan	15-4 (N/A)	Joe Bedenk	2 (last: 1957)	2nd (1957)	5–4
7	Western Michigan	MAC	24-7 (8-2, 0 GB)	Charlie Maher	3 (last: 1958)	2nd (1955)	7-6

1 title

2 context

Figure 6: Numeric symbol error.

title

context

caption

	Rank	Airport	Passengers handled	% Change 2011 / 12
0	1	Netherlands - Amsterdam	105,349	-
1	2	United Kingdom - Belfast-International	92,502	-
2	3	Portugal - Faro	71,676	-
3	4	Spain - Barcelona	66,565	-
4	5	Spain - Alicante	64,090	-
5	6	Spain - Malaga	59,175	-
6	7	Ireland - Dublin	35,524	-
7	8	Jersey - Jersey	35,169	1128
8	9	Ireland - Waterford	31,907	25
9	10	Spain - Palma de Mallorca	27,718	-

Figure 7: Delimiter drift error.

title

context

2002 MLS SuperDraft | Round three | | The 2002 Major League Soccer SuperDraft was held on February 10, 2002 in Lake Buena Vista, Florida. One month earlier, MLS had contracted two clubs: Tampa Bay Mutiny and Miami Fusion. In conjunction with the contraction, the league conducted both an Allocation Draft and a Dispersal Draft in which the top 2002 SuperDraft picks of Tampa Bay and Miami, as well as players from those clubs, were made available to remaining MLS clubs.passages:

caption

	Pick #	MLS team	Player	Position	Affiliation
0	26	Chicago Fire	Steve Totten	M	University of Virginia
1	27	Los Angeles Galaxy	Alejandro Moreno	F	UNC-Greensboro
2	28	Colorado Rapids	Bryn Ritchie	D	University of Washington
3	29	Colorado Rapids	Daniel Alvarez	M	Furman University
4	30	MetroStars	Sam Forko	D	University of Connecticut
5	31	D.C. United	Mohammed Fahim	F	Southern Methodist University
6	32	Kansas City Wizards	O'Neil Peart	F	Long Island Rough Riders
7	33	San Jose Earthquakes	Chris Roner	M	University of California
8	34	Colorado Rapids	Matt Moses	M	Stanford University
9	35	Columbus Crew	John Barry Nusum	F	Furman University
10	36	Chicago Fire	Dipsy Selolwane	F	St. Louis University
11	37	Kansas City Wizards	Chris Brunt	D	Southwest Missouri State
12	38	Los Angeles Galaxy	Cory Gibbs	D	Brown University
13	39	San Jose Earthquakes	Kevin Sakuda	D	Duke University

Figure 8: Cross-column linking error.

smallest in that field, the model incorrectly selects Connecticut. This reflects a structural formatting error in LaTeXthat disrupts numeric parsing rather than a misunderstanding of the concept of "wins."

D.1.3 Delimiter Drift

In Figure 7, the question "Which airport handled the most passengers?" has the gold answer "Netherlands - Amsterdam". The HTMLand LaTeXpredictions match this form, but the Markdown serialization normalizes the ASCII hyphen to a typographic en dash, yielding "Netherlands –Amsterdam". Under exact-match scoring, this punctuation substitution is marked incorrect despite being semantically identical. This error illustrates *delimiter drift*: Markdown's smart punctuation alters surface forms, so simple string equality fails unless we normalize symbols (e.g., mapping – → -) during evaluation.

D.1.4 Cross-column Linking

In Figure 8, the question asks, "What position was the player whose MLS team rebranded in November 2010, coinciding with its move to their home stadium?" The table lists draft picks from the 2002 Major League Soccer (MLS) SuperDraft, while the provided context (shown above) contains no explicit mention of any rebranding event. The correct answer is F (Forward), corresponding to O'Neil Peart of the Kansas City Wizards, a team that was later renamed *Sporting Kansas City* in 2010. The HTMLand LaTeXmodels correctly associate this external fact with the relevant table row, suggesting that they retrieved it from prior parametric knowledge rather than from the given context. In contrast, the Markdown model outputs D, corresponding to another player (Chris Brunt), due to weaker cross-column linkage. Markdown's flattened structure removes explicit cell boundaries between columns such as "MLS team" and "Player," increasing

	Historical population	Historical population	Historical population	Historical population
0	Census	Pop.		%±
1	1880	480		—
2	1890	633		31.9%
3	1900	727		14.8%
4	1910	650		−10.6%
5	1920	592		−8.9%
6	1930	597		0.8%
7	1940	761		27.5%
8	1950	932		22.5%
9	1960	1,085		16.4%
10	1970	1,214		11.9%
11	1980	1,401		15.4%
12	1990	1,349		−3.7%
13	2000	1,450		7.5%
14	2010	1,452		0.1%
15	Est. 2017	1,454		0.1%
16	U.S. Decennial Census	U.S. Decennial Census	U.S. Decennial Census	U.S. Decennial Census

Figure 9: Cell-to-row grounding error.

title

Britain Simons

context

caption

Film

	Year	Title	Role	Notes
0	2014	Babysitter's Black Book	Food Truck Boy	Lifetime Television
1	2015	Online Abduction (aka Cyber Case)	Ben	Lifetime Television
2	2016	Holidays	Handsome Boy	XYZ Films
3	2017	Ultraman X The Movie	Daichi Ozora (Voice)	Tsuburaya Productions
4	2017	Hickok	The Kid	Status Media & Entertainment

Figure 10: Relational composition error.

the chance of misbinding repeated entities. HTMLand LaTeXpreserve cell alignment, enabling more stable entity mapping across columns.

D.1.5 Cell-to-Row Grounding

In Figure 9, the model fails to generate the correct description for the highlighted cells `[[15,0],[15,1]]`, which correspond to the 2010 census row. The gold output is "Burton's population was 1,452 at the 2010 census." However, all formats (HTML, Markdown, and LaTeX) describe the subsequent row instead, producing sentences about the 2017 estimate (1,454). This error reflects a consistent *cell-to-row grounding* failure rather than a content misunderstanding: once the model selects the wrong row, it still generates a fluent and factually accurate sentence about that row. The issue arises because serialized table representations often include header rows or footnotes that shift the logical row indices by one. The model therefore grounds its generation on visual order instead of the true coordinate mapping.

D.1.6 Relational Composition

In Figure 10, the model is tasked with describing multiple highlighted cells `[[1,3],[2,0],[2,1],[2,3]]` from the filmography of Britain Simons. The correct output, "Britain Simons appeared in another Lifetime Television film (2015)", requires integrating information across rows and columns, linking the film title, year, and production company into a single relational statement. The HTMLmodel instead lists each cell independently (e.g., "The Notes cell for row 1 lists Lifetime Television; the Year cell for row 2 shows 2016..."), producing accurate fragments without composing them into one coherent sentence. This behavior reflects a *relational composition* error, where rigid structural markup enforces cell-level independence and discourages multi-cell synthesis. In contrast, Markdown and LaTeXprovide more linear textual representations that better support cross-cell aggregation and natural language fluency.

D.1.7 Row-ordering Error

In Figure 11, the question asks, "Which match was played the earliest in the tournament?" The table lists fixtures with dates such as 28 January, 1 February, and 6 February 1922. Several matches share the earliest date (28 January), so any of those fixtures would be acceptable. In our LaTeX serialization, the model selects a later match (Cardiff City vs. Southampton on 1 February), indicating a row-ordering misalignment: header lines, multirow or multicolumn constructs, and repeated rules can disrupt the perceived top-to-bottom sequence during tokenization, causing the temporal cue "earliest" to attach to a later entry.

D.2 SUC

We perform qualitative error analysis on the *SUC benchmark (SQA subset)* across three pipelines: GPT-4V (*VLM*), Qwen 2.5–3B (*VLM+Text*), and GPT-5 (*LLM*). All three systems exhibit consistent weaknesses in subtasks that require structural reasoning such as *row retrieval*, *table partition*, and *size detection*. Failures

Table 1:

	Tie no	Home team	Score	Away team	Date
0	1	Liverpool	0-1	West Bromwich Albion	28 January 1922
1	2	Preston North End	3-1	Newcastle United	28 January 1922
2	3	Southampton	1-1	Cardiff City	28 January 1922
3	Replay	Cardiff City	2-0	Southampton	1 February 1922
4	4	Leicester City	2-0	Fulham	28 January 1922
5	5	Nottingham Forest	3-0	Hull City	28 January 1922
6	6	Aston Villa	1-0	Luton Town	28 January 1922
7	7	Bolton Wanderers	1-3	Manchester City	28 January 1922
8	8	Swindon Town	0-1	Blackburn Rovers	28 January 1922
9	9	Tottenham Hotspur	1-0	Watford	28 January 1922
10	10	Barnsley	3-1	Oldham Athletic	28 January 1922
11	11	Northampton Town	2-2	Stoke	28 January 1922
12	Replay	Stoke	3-0	Northampton Town	1 February 1922
13	12	Brighton & Hove Albion	0-0	Huddersfield Town	28 January 1922
14	Replay	Huddersfield Town	2-0	Brighton & Hove Albion	1 February 1922
15	13	Bradford City	1-1	Notts County	28 January 1922
16	Replay	Notts County	0-0	Bradford City	1 February 1922
17	Replay	Notts County	1-0	Bradford City	6 February 1922
18	14	Crystal Palace	0-0	Millwall	28 January 1922
19	Replay	Millwall	2-0	Crystal Palace	1 February 1922
20	15	Southend United	0-1	Swansea Town	28 January 1922
21	16	Bradford Park Avenue	2-3	Arsenal	28 January 1922

1 title

2 context

Figure 11: Row ordering error

predominantly arise from errors in spatial alignment, header parsing, and over-generalization of table geometry from either visual or serialized inputs. Representative examples of these failure modes are summarized in Table 16.

Error Type	**Subtask**	**Ground Truth**	**Prediction**
Header-cell dropout (Visual)	Table Partition (GPT-4V)	Tour | $74,704	0 | $74,704
Row-span mis-localization	Row Retrieval (GPT-4V)	Connecticut | 20-1 (N/A) | J. Orlean Christian | 1 (last: 1957) | 5th (1957) | 1-2	36-8 | 36-8 | 24-5 | 24-5 | 38-11 | 38-11 | 22-4 | 22-4 | 15-4 | 15-4 | 24-7 | 24-7
Neighbor-row substitution	Row Retrieval (Qwen 2.5–3B)	American McGee Presents Akaneiro | 2010 | iPad | Spicy Pony | nan | Little Red Riding Hood in Japan	DexIQ | nan | nan | nan | Spicy Pony | nan | Puzzle Game
Null header parsing	Table First Cell (GPT-4V)	Liliana Barbulescu	nan
Column over-count	Size Detection (GPT-5)	8 | 6	8 | 7
Header index substitution	Table Partition (GPT-5)	Royal flush | 99.92%*	0 | 99.92%*

Table 16: Representative SUC benchmark failure cases across VLM, VLM+Text, and LLM pipelines on SQA.

D.2.1 Header-cell Dropout (Visual Segmentation Failure)

In the GPT-4V *table partition* subtask, the model predicted "0 | $74,704" instead of "Tour | $74,704". This indicates that the visual encoder failed to correctly recognize the first table cell, replacing it with an index token. The model likely focused on numeric spatial markers instead of OCR-extracted text, causing the header to collapse into a placeholder. Such failures are typical of purely visual models where attention maps overfit to grid positions rather than semantic content.

D.2.2 Row-span Mis-localization

For the *row retrieval* task, GPT-4V generated an entire concatenated column sequence ("36-8 | 36-8 | 24-5 | 24-5 ..."), while the correct row was "Connecticut | 20-1 (N/A) | J. Orlean Christian | 1 (last: 1957) | 5th (1957) | 1-2". This pattern shows that the model confused vertical and horizontal

token spans, producing column-wise aggregation rather than row isolation. It reveals a breakdown in visual-to-text alignment, where rows are not treated as atomic units but rather as feature stripes distributed across the table image.

D.2.3 Neighbor-row Substitution (Semantic Shortcut)

The *VLM+Text* model (Qwen 2.5–3B) retrieved the row "DexIQ" instead of the correct "American McGee Presents Akaneiro". Both rows share the developer "Spicy Pony", leading the model to rely on semantic similarity rather than spatial adjacency. This behavior reflects a *semantic-over-spatial* bias introduced by text linearization: in the serialized form, rows appear as flat textual sequences, erasing layout boundaries that encode which cells belong together.

D.2.4 Null Header Parsing

In the GPT-4V *table first cell* subtask, the model produced "nan" instead of "Liliana Barbulescu". This suggests that text extraction stage replaced the first textual entry with an empty token. Such cases highlight how missing-value placeholders can propagate through multimodal pipelines and be mistaken for actual content.

D.2.5 Column Over-count (Structural Over-generalization)

In the GPT-5 *size detection* task, the model predicted "8 | 7" instead of the correct "8 | 6". This overestimation indicates that uneven Markdown delimiters ("|") or extra HTMLtags were misread as additional columns. Without visual grounding, the model relies purely on string-level cues and thus hallucinates extra cells when markup spacing is inconsistent.

D.2.6 Header Index Substitution (Markup Confusion)

In the GPT-5 *table partition* subtask, the model replaced the true header "Royal flush" with a numeric "0". This occurs when serialized HTMLindices are interpreted as actual values, leading to the insertion of placeholder tokens at the start of a row. Such cases reveal that the LLM correctly understands structural intent (beginning and end of table) but fails to separate metadata tokens from meaningful cell content.