MMTABREAL: Real-World Benchmark for Multimodal Table Understanding

Anonymous ACL submission

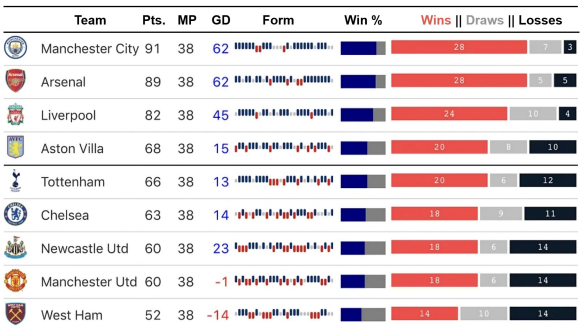
Abstract

Multimodal tables i.e. tabular layouts interleaved with charts, maps, icons, and color encodings are ubiquitous in real applications yet remain difficult for Multimodal Large Language Models (MLLMs). Despite advances in text and image understanding, systematic evaluation of table-centric multimodal reasoning is limited. We introduce MMTABREAL, a MultiModal Table Benchmark, human-curated suite of 500 real-world tables paired with 4,021 question–answer pairs. MMTABREAL spans four question types, five reasoning categories, and eight structural archetypes. Evaluations of state-of-the-art models reveal substantial gaps, especially in visual grounding, spatial alignment, and multi-step inference, with 20–40% performance drops relative to existing benchmarks. These results highlight the need for architectures that more tightly fuse vision with tabular structure and support explicit numeric/-logical operations. MMTABREAL is released for evaluation only, providing a rigorous, reproducible testbed that reflects the linguistic, structural, and reasoning complexity of real-world multimodal tables.

1 Introduction

Modern AI is shifting from unimodal perception to structure-aware multimodality, where models jointly read and reason over images embedded in tabular data in both structured and semi-structured (Lee et al., 2023; Cloutier and Ravasi, 2021). As systems move beyond single-channel inputs, seamless cross-modal integration becomes essential for human-level understanding across domains such as healthcare, finance, and education (Joshi, 2022).

Tables are a cornerstone of information representation, providing a two-dimensional scaffold for complex data in both structured and semi-structured forms (Shwartz-Ziv and Armon, 2022; Jiang et al., 2025). Modern tables are no longer text-only: they embed charts, images, color encodings, icons, and logos (Zheng et al., 2024). Unlike fixed-schema databases, this intuitive visual layout is easy for humans to parse but introduces distinct reasoning challenges: (i) header–cell hierarchies are implicit and must be inferred, and (ii) the semantics of adjacent cells are context-dependent (Shigarov, 2023).



	Team	Pts.	MP	GD	Form	Win %	Wins || Draws || Losses
	Manchester City	91	38	62			28 7 3
	Arsenal	89	38	62			28 5 5
	Liverpool	82	38	45			24 10 4
	Aston Villa	68	38	15			20 8 10
	Tottenham	66	38	13			20 6 12
	Chelsea	63	38	14			18 9 11
	Newcastle Utd	60	38	23			18 6 14
	Manchester Utd	60	38	-1			18 6 14
	West Ham	52	38	-14			14 10 14

Q1: Which team has the least number of draws? A1: Arsenal F.C.
Q2: What is the color of the badge of the club with 60 points and a negative goal difference? A2: Red
Q3: What is Chelsea's longest winning streak? A3: 3
Q4: Which club has the lowest goal difference for a team with a win percentage greater than 50? A4: Tottenham Hotspurs

Figure 1: A table of English Premier League standings along with accompanying questions and their respective answers from MMTABREAL.

Reasoning over such data, where textual, numerical, and visual cues are interwoven within a structured layout, is inherently challenging. For example, consider the question from Figure 1: *"Which club has the lowest goal difference among teams with a win percentage greater than 50%?"* Answering this requires multi-step, cross-modal reasoning: (1) detecting the visual bars in the *Win %* column to identify clubs exceeding 50%; (2) comparing numerical values in the *GD* (goal difference) column; (3) recognizing club badges, color encodings, and logos for correct team identification; (4) aligning these visual cues with the textual headers and cell

*Equal Contributor †Primary Superviser

contents to filter candidates; and (5) concluding that *Tottenham Hotspur* satisfies the condition with the lowest goal difference among the qualifying teams. Despite their ubiquity, multimodal tables remain underexplored in AI: most work targets generic vision-language or table-only tasks and overlooks the unique, structure-aware challenges of mixed-format tables (Vaishnav and Tammet, 2025). This gap motivates a central question: *Can today's AI systems, particularly multimodal large language models (MLLMs), effectively reason over complex multimodal tables?*

To address this question, we introduce MMTABREAL, a comprehensive, human-curated benchmark for question answering over *multimodal* tables. Unlike prior resources such as MMTabQA (Mathur et al., 2024) and MMTab (Zheng et al., 2024), which often rely on basic visuals or synthetic setups, it MMTABREAL spans multimodal web tables with rich, compositional cues: icons, logos, color encodings, and miniature charts paired with naturally phrased questions. MMTABREAL spans four question types (Explicit, Implicit, Answer-Mention, Visual-Based), five reasoning categories (Mathematical, Extrema Identification, Fact Verification, Vision-Based, Other), and eight structural archetypes (Single/Multiple Entity, Single/Multiple Chart, Maps, Maps+Charts with Entities, Visualizations).

Built from real-world tables and human-authored questions, it captures genuine linguistic variation and the practical reasoning challenges that arise in the real world. MMTABREAL *is for evaluation only not for training.* It provides a principled probe of current systems by testing whether models can fuse visual cues with tabular structure, align information spatially (headers, merged cells, row/-column layout), and perform numerical and logical operations. In doing so, MMTABREAL enables rigorous assessment of modern MLLMs on capabilities that remain beyond unimodal approaches. Our contribution are as follows:

- A human-curated benchmark, MMTABREAL, for multimodal table reasoning with real-world tables spanning **8** structural archetypes and **4K** naturally phrased QA pairs, featuring interleaved text, icons/logos, color encodings, and mini-charts.
- A standardized benchmarking of strong MLLMs shows **10–30%** performance drops relative to existing benchmarks across all baselines, indicating that human-curated, structurally authentic real-world tables QA nature of MMTABREAL.
- Fine-grained analyses across **table types**, **question types**, **reasoning skills** (visual detection, spatial alignment, numeric/logic operations), and **answer formats** reveal consistent models failure modes and actionable patterns.

2 Motivation

Why a new Multimodal TableQA benchmark? Existing tableQA datasets are limited in capturing the multimodal complexity of real-world data. MMTabQA (Mathur et al., 2024) and MMTab (Zheng et al., 2024) include basic visual elements like flags, other datasets such as MultimodalQA (Talmor et al., 2021), UniMMQA (Luo et al., 2023), and SPIQA (Yvinec et al., 2023) offer limited structural or visual diversity. Tables 1 and 2 summarize these gaps.

Dataset	#Test	Multimodal	Interleaved	TableImage	Hierarchical
MMTabQA	2,800	✓	✓	✓	✗
MMTab	49,000	✓	✗	✓	✓
MultimodalQA	3,660	✓	✗	✗	✗
UniMMQA	4,250	✓	✗	✗	✗
SPIQA	1,387	✓	✗	✗	✓
MMTABREAL	4,021	✓	✓	✓	✓

Table 1: Comparison of structural features in multimodal table QA datasets.

Dataset	Chart	Map	Visual	Flag	Char.	Loc.	Logo	Symbol
MMTabQA	✗	✗	✗	✓	✓	✗	✓	✓
MMTab	✗	✗	✗	✓	✓	✓	✓	✓
MultimodalQA	✗	✗	✗	✓	✓	✗	✓	✓
UniMMQA	✗	✗	✗	✓	✓	✓	✓	✓
SPIQA	✓	✗	✓	✗	✗	✗	✗	✗
MMTABREAL	✓	✓	✓	✓	✓	✓	✓	✓

Table 2: Comparison of multimodal QA datasets by visual content types.

MMTABREAL addresses these limitations by providing authentic multimodal tables that integrate diverse visual elements with complex tabular structures. MMTABREAL's design ensures that successful performance demands true multimodal understanding, making it an effective diagnostic tool for evaluating genuine reasoning capabilities in real-world scenarios.

Human-Curated vs. Synthetic Benchmarks Existing multimodal TableQA datasets such as MMTabQA (Mathur et al., 2024) and MMTab (Zheng et al., 2024) are synthetically constructed. Although synthetic generation enables scale, it often fails to capture the nuanced complexity of real-world multimodal tables. In contrast, human curation preserves authentic linguistic variation, realistic text–vision relationships, and genuine reasoning

challenges, while also mitigating issues common in AI-generated data—such as QA hallucinations (Ji et al., 2023; Bang et al., 2023) and spurious correlations (Zhao et al., 2024). Our focused scale further allows rigorous quality control that would be prohibitively costly at larger sizes, ensuring each table reflects authentic multimodal complexity. Crucially, human-curated benchmarks better match real use cases: meaning in multimodal tables often emerges from spatial layout and interdependent visual–textual cues. Modular pipelines that process text and images separately frequently discard this positional context, and real tables exhibit irregular layouts and functional visuals that resist synthetic replication (Gupta et al., 2022; Lee et al., 2023; Cloutier and Ravasi, 2021).

3 MMTabReal Benchmark

MMTabReal addresses key limitations of synthetic multimodal datasets by focusing on naturally occurring tabular formats and realistic QA needs.

3.1 MMTabReal Creation

Table Selection and Quality Control. We developed a custom pipeline to extract high-quality, structurally authentic multimodal tables from open-license, free-to-reuse sources, like Wikipedia/Wikimedia projects, U.S. federal open data (e.g., NASA, NOAA), and official open data portals such as data.europa.eu (CC BY 4.0) and UK OGL v3.0 sites. Multiple Selenium scripts were used for extraction, and tables available only as images (e.g., scanned documents or dashboards) were manually reconstructed to retain original formatting and ensure machine-readability. Minor fixes were applied to improve quality, including removing blurry images, correcting misaligned cells, and standardizing inconsistent formatting. All human faces originate from legally permissible public domain sources. Each table includes rich metadata describing structural properties, embedded image types, and visual content categories.

Annotation Process. Questions were designed to assess a model's ability to reason over both textual and visual elements in a table. They were created by NLP experts and cross-reviewed for correctness and consistency by peers and additional reviewers. Every question includes at least one image, either directly or via intermediate reasoning.

We prioritized comprehensive human annotation over semi-automated approaches using pre-trained large-language models, which often miss fine-grained visual details and nuanced reasoning. Automated generation also risks introducing model-specific biases, potentially favoring certain architectures. Our rigorous human curation ensures challenging, unbiased questions that provide an authentic evaluation of model capabilities.

Question Types. We follow the classification scheme of (Mathur et al., 2024), grouping questions into four types: (1) **Explicit Questions** directly reference an entity whose image is present in the table. These help evaluate a model's ability to directly link textual cues to visual content; (2) **Implicit Questions** involve an entity whose image is neither explicitly mentioned in the question nor in the answer but plays a crucial role in the intermediate reasoning process. These test a model's capacity for multi-step reasoning and inference over visual information ; (3) **Answer-Mention Questions** are characterized by answers containing an entity represented by an image in the table, while the question itself does not explicitly mention this entity. These help assess a model's ability to retrieve relevant visual entities even when they are not directly queried ; (4) **Visual-Based Questions** involve tasks that require direct analysis of visual aspects of images, such as color identification, shape recognition, and spatial relationships. They specifically assess a model's ability to perceive and interpret visual information.

Reasoning Types. Questions are categorized by the reasoning they require: (1) **Extrema Identification** asks for highest or lowest values within the table; (2) **Mathematical Questions** involve numerical operations and other quantitative computations; (3) **Fact Verification** questions involve retrieving information from the table to determine whether a given statement is correct; (4) **Vision-Based** questions involve analyzing visual elements, including shape, patterns; (5) **Others** encompasses a wide range of reasoning types not covered by other categories, including geography, common-sense and temporal reasoning, and general knowledge tasks.

3.2 MMTabReal Validation

MMTabReal underwent rigorous filtering and verification to ensure quality and reliability. Our process comprised three stages: automated filtering, manual quality control, and annotators validation.

Dataset Filtering: We implemented comprehensive filtering to remove low-quality content: (i)

Table filtering removed tables with excessive noise, missing critical data, or formatting issues. This initial process filtered 64 tables; (ii) *Content filtering* eliminated profanity, sensitive personal information, and inappropriate content using manual review; (iii) *Question editing* corrected ambiguous phrasing, unsolvable questions, and logical inconsistencies;

To ensure dataset quality and reliability, we implemented a comprehensive multi-stage annotation review and validation. All annotators used a standardized 3-point scale for question correctness evaluation: Score 0 for incorrect answers, 1 for partially correct answers, and 2 for fully correct answers. Our approach involved two phases of inter-annotator agreement analysis to assess question correctness.

Phase 1 - Internal Review (All Questions): Each question was reviewed and filtered by two other annotators who were not involved in creating that question, ensuring comprehensive quality control. Through this rigorous process, 11.26% of answers were corrected and 2.4% of tables were filtered to maintain dataset integrity. Table 3 reports Cohen's Kappa and agreement scores between the internal expert annotators on question correctness evaluation, demonstrating strong agreement across all question types.

Question Type	Cohen's κ	Percent Agreement
Explicit	0.86	96.2%
Implicit	0.80	93.1%
Answer-Mention	0.88	98.0%
Visual	0.77	90.9%
Overall Agreement	**0.82**	**96.3%**

Table 3: Inter-Annotator Agreement on Answer Correctness Evaluation

Phase 2 - External Validation (Subset): To address potential authorship bias, three external annotators unaffiliated with this research evaluated a subset of 20% of the questions (804 questions) with no prior knowledge of research objectives or gold standard answers.

Table 4 presents agreement scores between external annotators and gold answers, confirming the reliability of our work.

Question Type	Cohen's κ	Percent Agreement
Explicit	0.89	97.0%
Implicit	0.83	94.2%
Answer-Mention	0.91	98.8%
Visual	0.80	91.8%
Overall Agreement	**0.85**	**96.6%**

Table 4: Agreement Between External Reviewers and Gold Standard Evaluation

Additional details on Scoring and Filtering Guidelines, and Confusion Matrices for both agreements are provided in Appendix B.

3.3 MMTabReal Statistics

Dataset. Table 5 presents key statistics for MMTabReal, such as the total number of tables and questions, average structural dimensions, and the proportion of visual elements, illustrating the dataset's scale and multimodal richness.

Structural and Reasoning Diversity. Tables 6 and 7 show MMTabReal 's diversity: the first captures table layouts and answer formats, while the second presents question and reasoning type distributions, highlighting structural and cognitive richness.

Metric	Value
Total Tables	500
Total Questions	4,021
Avg. Images per Table	23.67
Avg. Rows per Table	19.49
Avg. Columns per Table	10.85
% Rows with Images	89.27
% Columns with Images	28.42

Table 5: MMTabReal Dataset Statistics

Table Types		Answer Types	
Type	**%**	**Type**	**%**
Single Entity	32.8	Singular Entity	56.5
Multiple Entity	26.6	Singular Number	23.9
Single Chart	10.8	Multiple Entities	13.5
Visualizations	9.6	Multiple Types	3.6
Entities+Maps	9.6	Multiple Numbers	1.9
Entities+Charts	4.2	Image Location	0.6
Multiple Charts	4.0		
Maps Only	2.4		
Total	**100.0**	**Total**	**100.0**

Table 6: Table and Answer Type Distributions

Question Types		Reasoning Types	
Type	**%**	**Type**	**%**
Explicit	55.5	Others	38.8
Implicit	18.4	Extrema	23.7
Visual	17.1	Mathematical	23.0
Answer-Mention	9.1	Visual	11.0
		Factual	3.5
Total	**100.0**	**Total**	**100.0**

Table 7: Question and Reasoning Type Distributions

Image Type Distribution. Table 8 shows the distribution of the types of images present in the dataset, demonstrating comprehensive coverage across visual domains relevant to real-world applications.

Image Type	% of Total	Image Type	% of Total
Human / Fictional Character	21.85	Chart	6.85
Flag / Coat of Arms / Seals	18.19	Location	5.34
Logo	15.73	Symbol	6.19
Map	9.01	Poster / Covers	5.76
Visualizations	2.02	Other	9.05
Total	**11,836**		

Table 8: Image Type Distribution

4 Experiments and Analysis

Using established multimodal evaluation practices, we benchmark on open and closed-source models.

4.1 Modeling Strategies

To evaluate model performance on our dataset, we adopt the five benchmarking strategies proposed by (Mathur et al., 2024): **(1) Missing Image Baseline** establishes a **lower performance bound** by removing all images from the table. Models must infer missing visual information solely from surrounding text, revealing their ability to reason under partial multimodal input. **(2) Entity Replaced Baseline** serves as an **upper performance bound**, manually replacing all images with precise textual descriptions. This setup measures reasoning ability under fully informative conditions, eliminating uncertainty from missing visuals. Image-to-text conversion follows a semi-automatic process: **entity images** and **maps** are replaced via Google Reverse Image Search, while **visualizations/charts** are described by multimodal LLMs. All outputs are human-verified. **(3) Image Captioning Baseline** transforms the multimodal task into a text-only setting by substituting each image with an automatically generated caption. Multimodal LLMs produce context-specific descriptions that are inserted into the tables, allowing analysis of how well models can extract and convey visual information through text. **(4) Table-as-Image Baseline** renders the entire table as an image, requiring models to interpret all information visually. We use Selenium to convert HTML tables while preserving structure and content, enabling assessment of models' ability to parse and reason over visually encoded table. **(5) Interleaved Baseline** preserves the original multimodal format, keeping images embedded within tables. This setting requires simultaneous reasoning over textual visual information, providing the comprehensive evaluation of models' capacity for multimodal integration and joint reasoning.

4.2 LLMs, Prompting, and Metric

LLMs: We evaluated multiple models across baselines: Gemini 1.5/2.0 (Team et al., 2024), GPT-4o Mini, Llama3-8b (Touvron et al., 2023), and Mixtral-8x7B (Jiang et al., 2024a) for text baselines. Vision-capable baselines additionally included InternVL2.5-8B (Kweon et al., 2023), Mantis-8B-Idefics2 (Jiang et al., 2024b), Phi-3.5-Vision (Abdin et al., 2024), Qwen 2.5-VL (Wang et al., 2024), Qwen 3-VL (Bai et al., 2025) and Table-LLaVA (Zheng et al., 2024). API calls were used for GPT and Gemini 1.5 and 2.0 inference. The remaining models were run locally from HuggingFace Transformers on an A100 GPU with 32 GB of memory. Parameters include a temperature of 0.2 and a maximum output length of 1024 tokens.

Prompts Strategies: We used 1-shot prompting for text baselines (Missing Image, Entity Replaced) and 0-shot for image baselines (Image Captioning, Table as Image, Interleaved). 1-shot helps text models by clarifying the task without overwhelming them (Sahoo et al., 2025), while examples can distract multimodal models (Ma et al., 2025), ensuring fair and optimal evaluation. Full prompts are in Appendix E.

Evaluation Metrics: We use standard metrics from prior work (Section 5) adapted to different answer types. **(a) Exact Match (EM)** checks if the prediction exactly matches the ground truth, used for single-number, single-entity, and image location answers. **(b) Substring Match (SS)** verifies if the prediction appears within the ground truth, allowing partial matches for multiple entities or numbers. **(c) F1-Score** computes the harmonic mean of precision and recall, suitable for multiple entities, numbers, and combinations.

4.3 Results and Analysis

Table 9, 10, and 15 (in Appendix), and 16 (in Appendix) analyses performance on MMTABREAL across several dimensions i.e., question types, table types, reasoning types and answer types, respectively.

Across modeling strategies: Missing Images performs poorly, highlighting the importance of visual data. Image Captioning shows the lowest gains, indicating MLLMs struggle with context-aware captioning. Table-as-Image and Interleaved perform similarly, though Table-as-Image is slightly lower, suggesting vision encoders struggle with dense single images and multiple-image integration remains challenging. Entity Replaced achieves the highest amongst the 4 types, but performs significantly lower than human baselines. Overall, these results point to architectural limitations rather than training deficiencies.

Effect of LLMs: Mixtral performs best in text-only scenarios with a low unknown rate (12%), showing that conservative uncertainty handling sup-

Model	Answer Mention			Explicit			Implicit			Visual Question		
	EM	SS	F1	EM	SS	F1	EM	SS	F1	EM	SS	F1
Missing Image Baseline												
Gemini 1.5 Flash	26.59	27.39	0.128	19.52	20.92	0.085	15.32	15.29	0.063	12.91	13.84	0.054
Gemini 2.0 Flash	27.98	30.15	0.089	19.31	21.03	0.075	14.12	14.77	0.052	17.60	18.51	0.064
GPT-4o Mini	38.99	38.40	0.294	33.97	36.24	0.251	24.14	25.71	0.143	27.00	27.33	0.163
Llama 3-8B	32.50	32.27	0.219	29.39	28.69	0.194	22.91	23.09	0.129	20.74	20.84	0.133
Mixtral	42.84	46.31	0.321	36.21	40.70	0.282	28.56	33.46	0.202	30.29	34.48	0.241
Entity Replaced Baseline												
Gemini 1.5 Flash	59.89	67.20	0.394	54.71	54.61	0.295	43.73	47.16	0.238	-	-	-
Gemini 2.0 Flash	59.50	62.46	0.293	59.93	60.04	0.300	39.71	41.26	0.177	-	-	-
GPT-4o Mini	68.14	70.38	0.538	65.99	69.67	0.496	50.59	52.73	0.340	-	-	-
Llama 3-8B	61.49	62.57	0.478	54.92	57.85	0.409	41.56	44.79	0.285	-	-	-
Mixtral	59.74	68.01	0.531	60.77	66.71	0.475	43.67	48.70	0.308	-	-	-
Image Captioning Baseline												
Gemini 1.5 Flash	29.70	30.79	0.224	30.12	32.74	0.219	18.91	19.45	0.126	21.44	24.04	0.156
Gemini 2.0 Flash	36.82	38.45	0.261	36.82	38.45	0.261	19.69	20.50	0.124	25.09	27.23	0.185
Table as an Image Baseline												
Gemini 1.5 Flash	38.39	36.22	0.178	30.16	31.30	0.148	25.14	27.52	0.113	25.66	27.80	0.103
Gemini 2.0 Flash	40.44	38.98	0.212	38.55	38.18	0.214	33.83	35.92	0.199	30.49	34.05	0.195
GPT-4o Mini	48.96	50.59	0.357	47.53	49.78	0.345	38.86	40.49	0.265	38.56	41.11	0.291
Intern-VL-2.5	19.55	40.26	0.199	18.55	38.53	0.176	16.42	36.90	0.153	14.47	38.63	0.162
Mantis	20.85	23.23	0.109	19.72	20.90	0.113	20.88	21.49	0.110	18.60	20.26	0.107
Phi-3.5	21.63	23.86	0.111	18.09	19.80	0.076	15.67	16.96	0.057	17.81	19.66	0.093
Qwen-2.5-VL	34.61	38.86	0.174	30.62	34.58	0.159	19.64	22.64	0.108	21.35	24.38	0.124
Qwen-3-VL	41.19	45.66	0.258	38.82	41.86	0.305	29.51	31.85	0.173	32.48	39.16	0.228
Table LLava-1.5-7B	10.30	11.43	0.062	12.68	14.49	0.063	15.77	16.52	0.060	10.95	11.30	0.050
Interleaved Baseline												
Gemini 1.5 Flash	34.38	35.24	0.247	31.55	31.52	0.210	20.33	20.47	0.119	26.29	25.65	0.175
Gemini 2.0 Flash	37.27	38.47	0.272	34.08	37.46	0.231	24.59	25.75	0.142	26.38	28.76	0.176
GPT-4o Mini	**47.74**	49.88	**0.376**	**46.92**	48.96	**0.348**	**36.41**	37.84	**0.260**	**40.39**	42.64	**0.303**
Mantis	24.76	26.45	0.156	24.37	26.57	0.150	24.92	26.58	0.113	20.70	23.12	0.126
Phi-3.5	20.85	23.72	0.120	21.63	23.61	0.114	23.83	26.85	0.134	17.71	18.95	0.100
Qwen-2.5-VL	35.66	**53.45**	0.271	30.35	**57.02**	0.258	17.95	**50.59**	0.146	23.04	**47.94**	0.200
Qwen-3-VL	41.19	45.66	0.358	38.82	41.86	0.305	29.51	31.85	0.173	32.48	39.16	0.294
Human Baseline	**78.4**	**82.1**	**0.76**	**84.2**	**87.3**	**0.81**	**75.8**	**80.6**	**0.73**	**79.9**	**83.7**	**0.78**

Table 9: Performance analysis across question types. EM: Exact Match, SS: Substring Match, F1: F1 Score. Human Baseline denotes a subset of the data.

ports reasoning without visual input. GPT-4o Mini exhibits higher uncertainty (45%) but excels in Entity Replaced scenarios, reflecting strength in symbolic reasoning when visual barriers are removed. Variations in "unknown" responses stem from differences in confidence calibration, risk tolerance, instruction-following, and training biases (Table 11). The vision-language alignment paradox appears in Qwen, which achieves the highest SS scores on Interleaved tasks but low EM, indicating strong content extraction but poor formatting. The resulting 18-point gap demonstrates that even advanced vision-language models can struggle with precise answer generation.

Model	Unknown %
Gemini 1.5 Flash	36.23%
Gemini 2.0 Flash	37.99%
GPT-4o Mini	45.96%
Llama 3-8B	41.54%
Mixtral-8x7B	12.13%

Table 11: Percentage of "Unknown" answers per model for Missing Image Baseline

Across Question Types: Answer-Mention questions perform best because they include direct lexical or visual references, letting models rely on simple cue matching without deeper reasoning. Explicit questions follow closely, as they still contain clear cues but require light compositional reasoning across short contexts. The narrow gap indicates that models handle shallow integration well when cues are explicit.

Implicit questions drop sharply because they demand multi-step inference and entity tracking without direct mentions, revealing weaknesses in maintaining contextual links once anchors are removed. Visual questions remain weak, as they rely on perceptual grounding rather than reasoning. While visual cues help stabilize attention, models still struggle to interpret visual details in context, making these tasks harder than those with explicit textual references.

Across Reasoning Types: Fact Verification achieves the highest performance, since these tasks resemble real-world factual checks with explicit cues. Mathematical reasoning performs the worst, highlighting persistent weaknesses in numerical logic. Even with visual context, models struggle to maintain consistency across calculation steps.

Extrema tasks perform slightly better, suggesting partial understanding of comparative or ordering relationships. Models can recognize relative patterns but often fail to generalize beyond simple comparisons. Vision-Based reasoning remains modest, as models struggle to link perceptual details with textual context. Visual cues offer some grounding,

Model	Single Entity			Multiple Entities			Single Chart			Multiple Charts		
	EM	SS	F1	EM	SS	F1	EM	SS	F1	EM	SS	F1
Table as an Image Baseline												
Gemini 1.5 Flash	39.36	39.83	0.22	38.41	39.07	0.20	40.78	41.55	0.23	31.39	33.46	0.22
Gemini 2.0 Flash	44.42	46.37	0.27	41.63	42.65	0.25	45.38	48.33	0.30	32.92	33.92	0.19
GPT-4o Mini	47.49	49.75	0.29	44.99	46.85	0.28	54.96	57.88	0.37	36.43	39.72	0.22
Intern-VL-2.5	33.08	45.69	0.25	35.03	53.24	0.27	34.55	56.89	0.25	35.58	50.58	0.25
Mantis-8B-Idefics2	30.04	30.80	0.16	32.77	34.51	0.19	44.98	46.31	0.24	37.55	47.14	0.24
Phi-3.5	31.75	32.72	0.18	26.82	28.15	0.16	36.09	37.87	0.20	31.62	32.28	0.20
Qwen-2.5-VL	39.65	41.64	0.23	36.53	39.51	0.23	36.58	38.96	0.21	35.63	36.81	0.23
Qwen-3-VL	43.20	45.30	0.26	40.25	42.80	0.25	48.50	51.20	0.32	35.80	38.10	0.21
Table_LLaVA	26.23	27.35	0.14	28.75	30.01	0.14	27.47	29.07	0.15	30.75	33.49	0.20
Interleaved Baseline												
Gemini 1.5 Flash	40.89	41.72	0.22	28.43	29.53	0.17	30.48	32.52	0.18	25.51	26.01	0.14
Gemini 2.0 Flash	40.96	42.42	0.22	32.77	33.99	0.19	40.77	43.63	0.22	33.33	34.67	0.20
GPT-4o Mini	45.83	47.26	0.28	43.14	44.48	0.27	52.99	55.61	0.34	38.56	41.39	0.27
Mantis-8B-Idefics2	34.80	36.59	0.19	34.24	36.19	0.18	40.66	41.27	0.22	42.38	42.99	0.23
Phi-3.5	39.75	41.65	0.21	38.28	42.32	0.24	44.23	46.71	0.25	33.28	34.08	0.20
Qwen-2.5-VL	29.65	61.37	0.19	29.06	64.65	0.18	30.10	76.22	0.18	30.53	59.01	0.17
Qwen-3-VL	42.50	44.20	0.25	39.80	41.50	0.24	47.20	49.80	0.30	35.30	37.60	0.24
Model	**Maps Only**			**Visualizations**			**Entities & Maps**			**Entities & Charts**		
Table as an Image Baseline												
Gemini 1.5 Flash	25.44	26.50	0.15	50.57	51.60	0.27	35.23	36.30	0.22	22.63	23.65	0.20
Gemini 2.0 Flash	28.27	29.30	0.18	55.56	56.60	0.30	38.66	39.70	0.25	24.64	25.65	0.22
GPT-4o Mini	27.12	28.15	0.17	54.40	55.45	0.29	40.03	41.05	0.27	23.35	24.38	0.21
Intern-VL-2.5	15.42	16.45	0.09	30.68	31.70	0.17	20.23	21.20	0.14	12.88	13.89	0.09
Mantis-8B-Idefics2	15.51	16.50	0.10	35.30	36.25	0.18	22.27	23.25	0.16	14.98	15.95	0.10
Phi-3.5	15.95	16.90	0.09	32.00	33.00	0.17	18.58	19.50	0.13	12.58	13.50	0.08
Qwen-2.5-VL	26.87	27.85	0.16	52.85	53.90	0.26	37.63	38.60	0.23	22.36	23.35	0.19
Qwen-3-VL	27.50	28.50	0.17	54.20	55.30	0.28	39.40	40.40	0.26	23.80	24.80	0.21
Table_LLaVA	12.28	13.28	0.08	28.71	29.75	0.15	19.73	20.70	0.12	10.84	11.85	0.07
Interleaved Baseline												
Gemini 1.5 Flash	23.44	24.50	0.13	48.57	49.60	0.25	33.23	34.30	0.20	20.63	21.65	0.18
Gemini 2.0 Flash	26.27	27.30	0.16	53.56	54.60	0.28	36.66	37.70	0.23	22.64	23.65	0.20
GPT-4o Mini	25.12	26.15	0.15	52.40	53.45	0.27	38.03	39.05	0.25	21.35	22.38	0.19
Mantis-8B-Idefics2	13.51	14.50	0.08	33.30	34.25	0.16	20.27	21.25	0.14	12.98	13.95	0.08
Phi-3.5	13.95	14.90	0.07	30.00	31.00	0.15	16.58	17.50	0.11	10.58	11.50	0.06
Qwen-2.5-VL	24.87	25.85	0.14	50.85	51.90	0.24	35.63	36.60	0.21	20.36	21.35	0.17
Qwen-3-VL	25.40	26.40	0.15	52.10	53.20	0.26	37.20	38.20	0.24	21.60	22.60	0.19

Table 10: Performance analysis across table types. EM: Exact Match, SS: Substring Match, F1: F1 Score

but current architectures lack the fine-grained spatial and semantic integration required for robust multimodal reasoning.

Across Table Types: Single-chart performance is highest, as real-world charts are designed for quick facts with consistent legends and scales, making them easy for humans and MLLMs to interpret. Performance declines with multiple charts or chart-plus-entity setups, since the model must align scales and legends and combine information across visuals, requiring more reasoning and structural understanding.

The gap between single and multiple entity inputs is small. MLLMs, trained on many images with diverse entities, can recognize and label them easily. The main challenge lies in linking visual and textual information in structured layouts, not the number of entities.

Across Answer Types: Image Location and Multiple Types perform poorly, as models struggle to integrate spatial, numeric, and entity data, which requires cross-modal reasoning that current architectures handle only partially.

Entity-based answers perform best, with Single and Multiple Entity types showing strong results. Entities are frequent in training data and easy to recognize, so models can rely on surface-level visual and lexical cues. Performance is similar for single and multiple entities since identification remains straightforward even with multiple targets.

Number-based answers are intermediate. Single Number tasks are moderately challenging, showing partial success in numeric retrieval. Multiple-number tasks perform better than expected because they do not require symbolic computation, highlighting that models struggle with mathematical reasoning but can retrieve numeric information without operations.

4.4 Our Key Findings:

Our analysis reveals several specific limitations in current MLLMs. **1) Table and structured data understanding is poor**, as models struggle to interpret chart layouts, align scales, and integrate multiple types of information, resulting in very low performance on Image Location and Multiple Types answers. **2) Explicit cue recognition is strong**, with models performing better on tasks where textual or visual references are directly present, reflecting a reliance on surface-level matching rather than deeper reasoning. **3) Vision encoders are limited**, as models fail to capture subtle spatial or perceptual

details in images, constraining their ability to reason over visual content. **4) Performance declines as table complexity increases**, with Single Chart tables outperforming Multiple Chart tables and Entities & Charts showing the lowest performance, highlighting the added reasoning challenges.

We also conducted detailed human evaluation and error analysis on a MMTABREAL subset, see Appendix D.

5 Comparison with Related Works

As shown in Tables 1 and 2 in Section 2, existing datasets fall short of capturing the complexity of multimodal tables across key dimensions:

(1) *Limited structural coverage:* Text-only datasets such as TaT-QA (Zhu et al., 2021), KET-QA (Hu et al., 2024), FinQA, TabMWP (Lu et al., 2023b) (Chen et al., 2021), and DynaQA (Lu et al., 2022) omit visual information entirely, while early multimodal efforts—MMCoQA (Li et al., 2022), MMTab (Zheng et al., 2024), and InfoSeek (Chen et al., 2023)—include visuals only in a limited and loosely aligned manner. In contrast, visual reasoning datasets like ChartQA (Masry et al., 2022), mChartQA (Wei et al., 2024), and MMC (Liu et al., 2024) focus on images but lack tabular structure;

(2) *Disjoint modalities:* Datasets such as MMQA (Talmor et al., 2021), UniMMQA (Luo et al., 2023), and SPIQA (Yvinec et al., 2023) treat text, tables, and images as separate components rather than cohesive multimodal entities, failing to capture the natural interplay of visual and textual information in real-world documents; and

(3) *Synthetic limitations:* MMTabQA (Mathur et al., 2024) moves toward multimodal integration by extending text-based datasets (e.g., FetaQA (Nan et al., 2022), HybridQA (Chen et al., 2020), WikiSQL (Zhong et al., 2017), Open-WikiTable (Kweon et al., 2023)) with generated visuals, but its synthetic construction and Wikipedia-only origin limit domain diversity and fail to reproduce authentic visual and structural complexity.

MMTABREAL **vs. MMTabQA** Among existing datasets, MMTabQA is the closest to MMTABREAL in scope and modality. However, MMTABREAL poses substantially greater challenges, as shown by uniformly lower baseline performance in Table 12. The largest drops occur in the Missing Image and Image Captioning settings, indicating that MMTabQA models could more easily infer missing entities under simpler conditions. Even the newer Gemini 2.0 Flash lags behind Gemini 1.5 Flash on our benchmark, underscoring MMTABREAL 's increased difficulty.

Baseline Setting	MMTabQA	MMTABREAL
Missing Image (GPT 4o vs Mistral)	62.72	38.85
Entity Replaced (GPT 4o vs GPT 4o - mini)	78.54	65.95
Image Captioning (Gemini 1.5 vs 2.0 Flash)	52.33	33.26
Table as Image (GPT 4o vs GPT 4o - mini)	58.58	46.71
Interleaved (GPT 4o vs Qwen 2.5 VL)	61.08	53.45

Table 12: Baseline comparison on substring match accuracy. MMTABREAL results show consistently lower performance across settings.

The Table-as-Image baseline also declines sharply, reflecting the added complexity of larger tables and multiple images per instance. While the Interleaved baseline shows a smaller decrease, Qwen 2.5 VL 7B still outperforms GPT-4o on several external benchmarks (e.g., DocVQA (Mathew et al., 2021), CC-OCR (Yang et al., 2024), MathVista (Lu et al., 2023a), AITZ (Zhang et al., 2024), ScreenSpot (Li et al., 2025)), highlighting the competitiveness of current multimodal models. Overall, the consistent performance degradation confirms that MMTABREAL constitutes a more challenging diagnostic benchmark for multimodal reasoning.

6 Conclusion

In conclusion, we introduce MMTABREAL, a benchmark designed to evaluate multimodal reasoning over real-world tables. Our experiments reveal substantial gaps in current models, with exact match scores ranging from 21.6–48.1% and near-zero performance on Image Location and spatial reasoning tasks. Models struggle with visual-spatial reasoning, complex table structures, and tasks requiring genuine multimodal understanding, often relying on surface-level cues. MMTABREAL 's diverse visual and structural coverage makes it a critical benchmark for diagnosing and emphasizing the need for better integration of textual and visual information in multimodal models.

From our analysis, we suggest future avenues in multimodal table reasoning: (a) **Visual-Spatial Architectures:** Develop specialized attention mechanisms for precise spatial localization to address poor structural reasoning. (b) **Compositional Reasoning**: Develop systems capable of understanding and integrating information to perform multi-step inference and structured reasoning, going beyond simple pattern matching. (c) **Scalable Dataset Generation:** Explore automated methods that preserve authentic visual-structural complexity while expanding beyond manual curation constraints.

Limitations

Our primary limitation is scalability due to the need for expert-driven curation. Manual creation and verification of multimodal tables ensure high-quality, semantically coherent data but make large-scale expansion challenging compared to automated generation methods. However, automated approaches would likely overlook the nuanced visual-structural relationships that define multimodal table reasoning, which are central to MMTABREAL 's strength. Thus, while scalability remains a constraint, it is a deliberate trade-off to preserve interpretability and evaluative precision.

A secondary limitation concerns test set size. While 500 tables may appear limited at first glance, MMTABREAL is designed explicitly as an evaluation benchmark rather than a training corpus. Its scale is consistent with the test set sizes of other multimodal table QA datasets, as shown in Table 1. Moreover, MMTABREAL includes a wider variety of image types, and more complex hierarchical structures, making it a richer and more diagnostic benchmark despite its focused scope.

Ethical Statement

As the authors of this work, we confirm that our research and publication comply with the highest ethical standards. We used AI tools to assist with the writing process. The dataset used in this study is intended solely for academic purposes and should not be used for other purposes, allowing the scientific community to verify and build upon our work. All images included in the dataset are sourced from the public domain, have undergone filtering to remove sensitive content, and are free to use; even face images are publicly available and cleared for research purposes. The claims presented in this paper are consistent with the results of our experiments. To ensure reproducibility, we have provided detailed information on the prompting methods, models, and annotations used. All three reviewers involved in the inter-annotator agreement process participated voluntarily and provided informed consent.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *Preprint*, arXiv:2302.04023.

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

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *ArXiv*, abs/2302.11713.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charlotte Cloutier and Davide Ravasi. 2021. Using tables to enhance trustworthiness in qualitative research. *Strategic organization*, 19(1):113–133.

Abhishek Gupta, Shreshta Rajakumar Deshpande, and Marcello Canova. 2022. An algorithm to warm start perturbed (wasp) constrained dynamic programs. *IEEE Open Journal of Control Systems*, 1:1–14.

Mengkang Hu, Haoyu Dong, Ping Luo, Shi Han, and Dongmei Zhang. 2024. KET-QA: A dataset for knowledge enhanced table question answering. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9705–9719, Torino, Italia. ELRA and ICCL.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas,

Emma Bou Hanna, Florian Bressand, and 1 others. 2024a. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024b. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Jun-Peng Jiang, Si-Yang Liu, Hao-Run Cai, Qile Zhou, and Han-Jia Ye. 2025. Representation learning for tabular data: A comprehensive survey. *arXiv preprint arXiv:2504.16109*.

Abhinav Joshi. 2022. Multimodal representation learning for real-world applications. ICMI '22, page 717–723, New York, NY, USA. Association for Computing Machinery.

Sunjun Kweon, Yeonsu Kwon, Seonhee Cho, Yohan Jo, and Edward Choi. 2023. Open-WikiTable : Dataset for open domain question answering with complex reasoning over table. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8285–8297, Toronto, Canada. Association for Computational Linguistics.

Gyeong-Geon Lee, Lehong Shi, Ehsan Latif, Yizhu Gao, Arne Bewersdorff, Matthew Nyaaba, Shuchen Guo, Zihao Wu, Zhengliang Liu, Hui Wang, and 1 others. 2023. Multimodality of ai for education: Towards artificial general intelligence. *arXiv preprint arXiv:2312.06037*.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. 2025. Screenspot-pro: Gui grounding for professional high-resolution computer use. *arXiv preprint arXiv:2504.07981*.

Yongqi Li, Wenjie Li, and Liqiang Nie. 2022. MMCoQA: Conversational question answering over text, tables, and images. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4220–4231, Dublin, Ireland. Association for Computational Linguistics.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1287–1310, Mexico City, Mexico. Association for Computational Linguistics.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023a. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *Preprint*, arXiv:2209.14610.

Haohao Luo, Ying Shen, and Yang Deng. 2023. Unifying text, tables, and images for multimodal question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9355–9367, Singapore. Association for Computational Linguistics.

Xinbei Ma, Yiting Wang, Yao Yao, Tongxin Yuan, Aston Zhang, Zhuosheng Zhang, and Hai Zhao. 2025. Caution for the environment: Multimodal LLM agents are susceptible to environmental distractions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22324–22339, Vienna, Austria. Association for Computational Linguistics.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Suyash Vardhan Mathur, Jainit Sushil Bafna, Kunal Kartik, Harshita Khandelwal, Manish Shrivastava, Vivek Gupta, Mohit Bansal, and Dan Roth. 2024. Knowledge-aware reasoning over multimodal semi-structured tables. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14054–14073, Miami, Florida, USA. Association for Computational Linguistics.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. 2022. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2025. A systematic survey of prompt engineering in large language models: Techniques and applications. *Preprint*, arXiv:2402.07927.

Alexey Shigarov. 2023. Table understanding: Problem overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(1):e1482.

Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Mohit Vaishnav and Tanel Tammet. 2025. Cognitive paradigms for evaluating vlms on visual reasoning task. *arXiv preprint arXiv:2501.13620*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jingxuan Wei, Nan Xu, Guiyong Chang, Yin Luo, BiHui Yu, and Ruifeng Guo. 2024. mchartqa: A universal benchmark for multimodal chart question answer based on vision-language alignment and reasoning. *arXiv preprint arXiv:2404.01548*.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Yuliang Liu, and 1 others. 2024. Cc-ocr: A comprehensive and challenging ocr benchmark for evaluating large multimodal models in literacy. *arXiv preprint arXiv:2412.02210*.

Edouard Yvinec, Arnaud Dapogny, Matthieu Cord, and Kevin Bailly. 2023. Spiq: Data-free per-channel static input quantization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3869–3878.

Jiwen Zhang, Jihao Wu, Yihua Teng, Minghui Liao, Nuo Xu, Xiao Xiao, Zhongyu Wei, and Duyu Tang. 2024. Android in the zoo: Chain-of-action-thought for gui agents. *arXiv preprint arXiv:2403.02713*.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. 2024. What makes unlearning hard and what to do about it. *Preprint*, arXiv:2406.01257.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *arXiv preprint arXiv:2105.07624*.

A Dataset and Question Types Examples

This section provides representative examples from each of the eight table structure types identified in Table 6, demonstrating the diversity and complexity of multimodal reasoning challenges across different structural configurations.

A.1 Single Entity

RK	DEFENSE	G	DRIVES	STOP RATE	PTS/DRIVE
1	Minnesota	4	40	85%	0.85
2	Michigan	6	72	83.3%	0.94
3	San Jose State	4	46	82.6%	1.02
4	Illinois	5	65	81.5%	0.74
5	Penn State	5	64	81.2%	1.05
6	Alabama	6	84	81%	0.89
7	James Madison	4	56	80.4%	1.21
8	Iowa State	5	60	80%	1.20
9	Iowa	5	59	79.7%	0.95
10	Ohio State	6	73	78.1%	1.19

Figure 2: An example of single entity multimodal table in the sports domain with only one type of entity (Image of Logo+Text).

Figure 2 is a collegiate football defensive performance table that ranks the top ten teams based on their Stop Rate, the percentage of opponent drives that end without a score. The table includes five columns detailing performance metrics.

Example Questions:

Q: What is the Stop Rate for the Fighting Illini?
A: 81.5% **Type:** Explicit.

Q: What is the difference in DRIVES faced between the highest-ranked team that played 4 Games and the highest-ranked team that played 6 Games?
A: 32 **Type:** Implicit.

Q: Which team has an allowed PTS/DRIVE that is exactly 1.21?
A: James Madison **Type:** Answer-Mention.

Q: How many teams have at least one letter from their initials in their logos?
A: 7 **Type:** Visual.

A.2 Multiple Entity

Figure 3 is a Formula 1 Race Results Summary Table detailing the performance of multiple drivers across three consecutive racing seasons. The table is organized by Year, and each year begins with a header row displaying the Race Location Flags for the first three races of that season. The subsequent rows for each year list the Driver, their Nationality, and their finishing position in the three respective races.

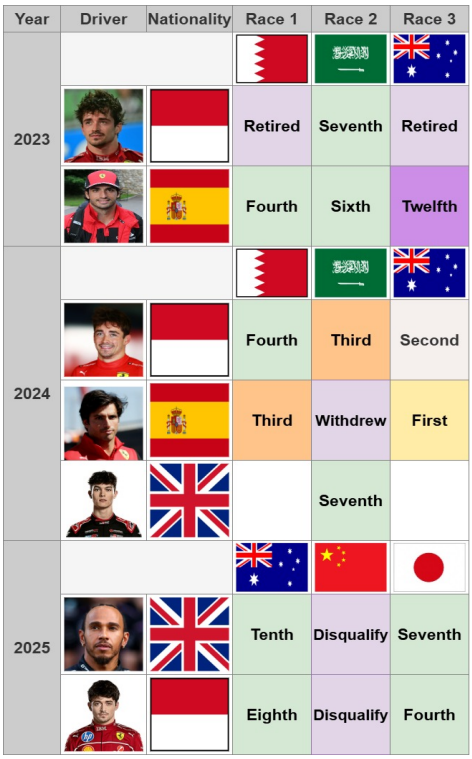


Figure 3: An example of multiple entity multimodal table in the motorsports domain with 2 types of entities (Image of flags + humans).

Example Questions:

Q: How many British drivers finished in the top 5 in Australia?
A: 0 **Type:** Explicit.

Q: What was the driver's position at the same venue one year after he finished 12th?
A: First **Type:** Implicit.

Q: Which driver replaced the driver whose national flag contains yellow, for one race?
A: Oliver Bearman **Type:** Answer-Mention.

Q: How many stars appear in this table?
A: 23 **Type:** Visual.

A.3 Single Chart

Figure 2 is a Product Sales Performance Dashboard that displays the performance of six products (Amarilla, Carretera, Montana, Paseo, Velo, VTT) and a Total row across two metrics: Sum of Gross Sales, which shows each product's percentage contribution to total sales, and a line chart labeled Sum of Gross Sales by Month Number, which visualizes the monthly sales trend for each product.

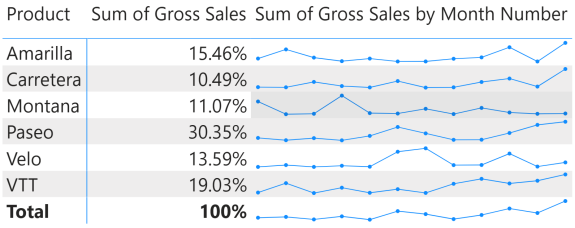


Figure 4: An example of a multimodal table with 1 type of chart(sparkline)

Example Questions:

Q: What is the difference in Overall gross sales between the product whose sales rose the most after March vs dropped the most after March?
A: 0.58% **Type:** Implicit.

Q: Which Product has the most number of drops followed by a peak?
A: VTT **Type:** Visual.

A.4 Multiple Charts

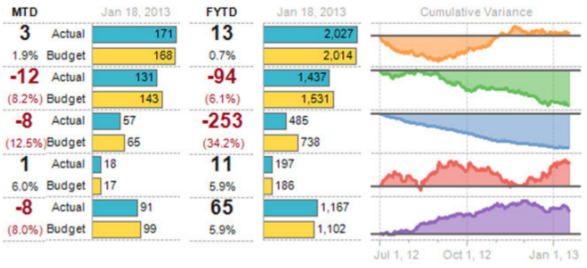


Figure 5: An example of a multimodal table in a Financial Context with multiple chart types (bar chart + area chart)

Figure 5 presents a multimodal financial dashboard displaying Monthly-to-Date (MTD) and Fiscal-Year-to-Date (FYTD) comparisons between actual and budgeted values across five performance metrics. It also includes visual representations of cumulative variance trends.

Example Questions:

Q: What is the actual MTD for the third metric on January 18, 2013?
A: 57 **Type:** Explicit.

Q: How many FYTD units were sold in total?
A: 5313 **Type:** Implicit.

Q: Which metric reported the greatest increase?
A: FYTD **Type:** Answer-Mention.

Q: What colored variance graph shows a steady decline over time?
A: Blue **Type:** Visual.

A.5 Visualizations

Figure 6 is a table detailing the design and components of four chemical compounds or drug candidates, identified by an ID number. The table presents a common Core Fragment structure for all compounds containing three variable attachment points. The subsequent three columns (R1, R2, R3) display the specific chemical groups or fragments intended to be attached at each corresponding position on the core structure.

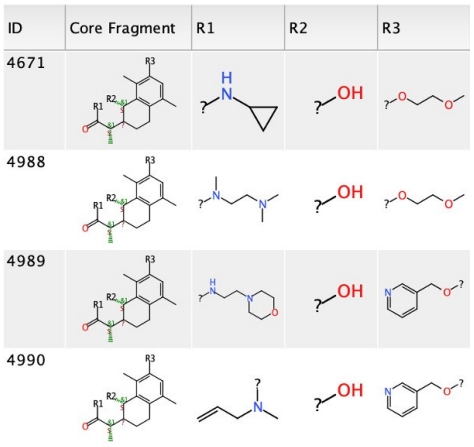


Figure 6: An example of multimodal table in the chemistry domain with scientific visualizations.

Example Questions:

Q: Which ID numbers have a benzene bases R3?
A: [4989, 4990] **Type:** Explicit.

Q: How many Nitrogen atoms exist in R1 and R3 combined?
A: 8 **Type:** Implicit.

Q: What is the name of the common ion in R2?
A: Hydroxide Ion **Type:** Answer-Mention.

Q: How many Hydrogen atoms exist in the R1 of 4988?
A: 13 **Type:** Visual.

A.6 Entities+Maps

Figure 7 is a table detailing four former franchises from the Indian Premier League (IPL). The table presents four columns: the team's logo and name, the home City, the State represented by a map of India with the relevant state highlighted in red, and the year the team made its league Debut.

Team	City	State	Debut
	Hyderabad		2008
	Kochi		2011
	Pune		2011
	Pune		2016

Figure 7: An example of a multimodal table in the sports domain with maps and entities. (Image of logos + highlighted map)

Example Questions:

Q: In which year did Pune Warriors India debut?
A: 2011 **Type:** Explicit.

Q: Which state does the team with the purple logo play in?
A: Maharashtra **Type:** Implicit.

Q: Which team from South India debuted in 2011?
A: Kochi Tuskers Kerala **Type:** Answer-Mention.

Q: How many teams have animals in their logo?
A: 3 **Type:** Visual.

A.7 Entities+Charts

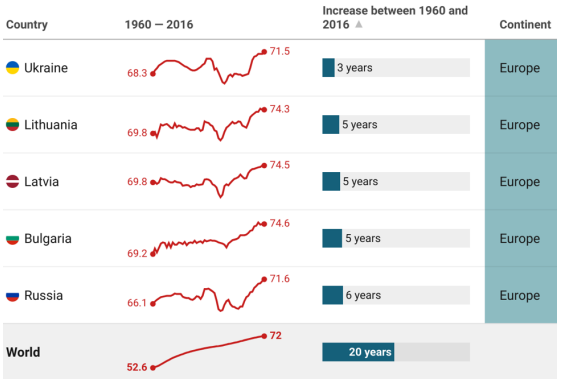


Figure 8: An example of a multimodal table in the Social Sciences Domain with charts and entities (Images of country flags, sparkline chart, and bar charts).

Figure 8 is a data visualization comparing the change in life expectancy across five Eastern European countries and the global average between 1960 and 2016. The table features four columns: Country, a line chart showing the trend, the total Increase between 1960 and 2016, and the Continent.

Example Questions:

Q: Has the score of Ukraine ever dropped below its starting point?
A: Yes **Type:** Explicit.

Q: By how many percent did the global data increase throughout the years?
A: 36.88% **Type:** Implicit.

Q: Which Country had the highest peak?
A: Bulgaria **Type:** Answer-Mention.

Q: By how many points did the country whose flag does not contain any red grow?
A: 3.2 **Type:** Visual.

A.8 Maps Only

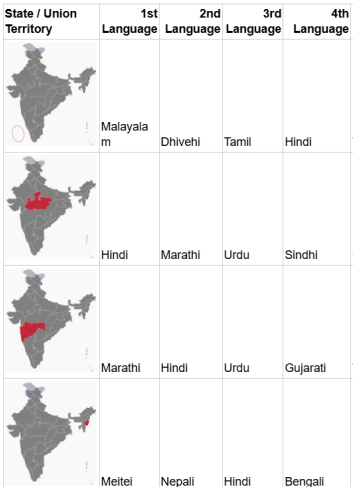


Figure 9: An example of a multimodal table in the Linguistics Domain with maps

Figure 9 is a table that displays the four most spoken languages in four different States or Union Territories of India. The table features five columns: an image column showing a map of India with the relevant State / Union Territory highlighted in red, followed by four columns listing the 1st, 2nd, 3rd, and 4th most spoken Language in that region.

Example Questions:

Q: What is Lakshadweep's third language?
A: Tamil **Type:** Explicit.

Q: What is the first language of the region that borders Assam?
A: Meitei **Type:** Implicit.

Q: Name all states/union territories that speak Hindi?
A: [Lakshadweep, Madhya Pradesh, Maharashtra, Manipur] **Type:** Answer-Mention.

Q: How many states are completely landlocked?
A: 2 **Type:** Visual.

B Inter-Annotator Guidelines

To ensure dataset quality and reliability, we implemented a comprehensive multi-stage annotation process with independent review and validation. Our approach involved two phases of inter-annotator agreement analysis to assess question correctness across all 4,021 questions.

Question Correctness Evaluation Criteria: All annotators used a standardized 3-point scale for question correctness evaluation with detailed guidelines to ensure consistency:

- **Score 0 (Incorrect):** Questions that contain factual errors, ambiguous phrasing that prevents accurate answering, or answers that cannot be derived from the provided table content. This includes questions with incorrect references to table elements, logical inconsistencies, or answers that contradict visual or textual information in the table.
- **Score 1 (Partially Correct):** Questions that are generally well-formed but have minor issues such as slight ambiguity in phrasing, answers that are approximately correct but lack precision, or questions that could benefit from clearer wording. These questions are answerable but may require interpretation or have multiple plausible answers.
- **Score 2 (Fully Correct):** Questions that are clearly formulated, unambiguous, and have definitive answers that can be accurately derived from the table content. These questions demonstrate appropriate difficulty level, require genuine multimodal reasoning, and align perfectly with the provided visual and textual information.

Table Filtering Criteria: Annotators evaluated tables using comprehensive guidelines to ensure dataset quality and appropriateness for multimodal reasoning evaluation:

- **Structural Coherence Requirements:** Tables must demonstrate logical organization with clear relationships between textual and visual components. The structure should support meaningful multimodal reasoning tasks, with appropriate alignment between headers, data cells, and embedded visual elements.
- **Content Complexity Thresholds:** Tables must contain sufficient complexity to warrant multimodal analysis, including diverse visual elements (charts, maps, images), hierarchical data organization, and content that requires integration of both visual and textual information for comprehension.
- **Reasoning Appropriateness:** Tables should enable authentic multimodal reasoning scenarios where visual elements are essential for answering questions, rather than serving merely decorative purposes. The content must support various question types including fact verification, mathematical calculation, extrema identification, and visual-based inference.
- **Profanity and Sensitivity Compliance:** Tables must not contain or depict any profane, explicit, hateful, or sensitive content, including personally identifiable information (PII) or imagery that violates ethical, cultural, or privacy standards. All content should be appropriate for general academic and research use.
- **Data Source and Open-Domain Validity:** All table data and visual content should originate from credible, open-domain, and publicly accessible sources. Datasets must be free of copyright or usage restrictions that prevent open dissemination, ensuring reproducibility and transparency in multimodal research.

Cohen's Kappa (κ) was computed on independent judgments to measure inherent agreement across both evaluation criteria.

Phase 1 - Internal Review: Table 3 reports Cohen's Kappa and Table 13 provides the confusion matrix for those scores.

Annotator A	Annotator B 0	1	2	Total
0	156	18	4	178
1	22	298	45	365
2	8	51	3419	3478
Total	186	367	3468	4021

Table 13: Confusion Matrix: Internal Expert Annotators (Phase 1)

Phase 2 - External Validation (Subset): Table 4 reports Cohen's Kappa and Table 14 provides the confusion matrix for those scores.

External Annotators	Gold Standard 0	1	2	Total
0	32	3	1	36
1	4	71	8	83
2	2	9	674	685
Total	38	83	683	804

Table 14: Confusion Matrix: External Annotators vs Gold Standard (Phase 2)

C Additional Result Tables

This section presents detailed performance results across all reasoning types and answer types discussed in Section 4.

Model	Fact Verification			Mathematical			Extrema			Vision Based		
	EM	SS	F1	EM	SS	F1	EM	SS	F1	EM	SS	F1
Missing Image Baseline												
Gemini 1.5 Flash	28.86	29.85	0.179	16.42	17.52	0.059	16.73	17.47	0.070	5.23	5.52	0.030
Gemini 2.0 Flash	27.62	27.61	0.099	15.96	16.35	0.049	15.80	16.20	0.059	11.00	11.04	0.032
GPT-4o Mini	44.75	47.64	0.361	23.80	23.29	0.124	27.45	27.07	0.146	24.09	25.64	0.177
Llama 3-8B	38.48	40.60	0.258	20.25	22.18	0.095	18.15	20.17	0.115	23.01	23.85	0.168
Mixtral	47.43	52.33	0.429	24.66	30.91	0.174	27.29	32.58	0.184	29.07	33.38	0.235
Entity Replaced Baseline												
Gemini 1.5 Flash	68.18	72.65	0.322	40.10	42.48	0.210	57.01	60.21	0.324	-	-	-
Gemini 2.0 Flash	71.69	73.28	0.333	42.11	41.20	0.191	64.80	65.86	0.347	-	-	-
GPT-4o Mini	79.45	75.11	0.628	39.00	43.92	0.255	54.06	57.00	0.357	-	-	-
Llama 3-8B	70.45	74.84	0.586	32.15	37.40	0.205	41.02	44.32	0.270	-	-	-
Mixtral	80.17	83.34	0.643	34.05	39.97	0.237	44.16	49.18	0.317	-	-	-
Image Captioning Baseline												
Gemini 1.5 Flash	47.90	49.29	0.369	13.68	15.38	0.074	25.13	27.47	0.165	22.72	23.76	0.186
Gemini 2.0 Flash	48.14	51.87	0.358	19.60	21.43	0.092	28.48	33.31	0.179	29.15	30.22	0.235
Table as an Image Baseline												
Gemini 1.5 Flash	47.17	45.77	0.207	23.42	25.33	0.095	21.30	23.01	0.113	30.40	32.38	0.094
Gemini 2.0 Flash	45.75	46.04	0.254	29.57	32.16	0.174	29.92	30.27	0.179	29.11	30.72	0.163
GPT-4o Mini	63.34	64.06	0.490	34.00	36.52	0.214	41.62	43.86	0.275	38.50	41.45	0.306
InternVL 2.5-8B	19.94	46.19	0.222	15.80	38.13	0.154	15.94	33.66	0.148	11.76	38.97	0.141
Mantis-8B-Idefics2	23.28	27.00	0.155	19.63	22.24	0.084	15.54	16.20	0.071	19.07	22.51	0.118
Phi-3.5-vision-instruct	32.16	35.79	0.238	13.87	16.18	0.052	13.73	15.01	0.063	15.24	17.70	0.075
Qwen-2.5-VL	53.12	55.19	0.265	21.45	25.08	0.113	27.90	31.39	0.158	22.54	24.40	0.102
Qwen-3-VL	58.20	60.50	0.380	28.50	31.20	0.165	35.80	38.40	0.220	31.80	34.20	0.215
Table-llava-1.5-7b-hf	24.57	27.06	0.133	14.57	15.51	0.038	8.42	9.54	0.031	9.88	10.93	0.049
Interleaved Baseline												
Gemini 1.5 Flash	41.37	41.14	0.283	17.96	18.00	0.079	27.78	27.21	0.155	26.81	27.28	0.213
Gemini 2.0 Flash	43.66	48.04	0.332	20.08	21.56	0.092	30.18	32.88	0.153	31.51	32.62	0.232
GPT-4o Mini	61.40	61.91	0.479	28.84	29.66	0.174	38.78	40.97	0.265	39.69	41.87	0.316
Mantis-8B-Idefics2	29.13	32.50	0.190	20.01	21.94	0.082	18.83	20.40	0.097	27.06	26.81	0.144
Phi-3.5-vision-instruct	25.67	28.31	0.168	21.98	24.72	0.107	16.39	17.39	0.093	17.33	19.14	0.106
Qwen-2.5-VL	43.03	62.76	0.381	15.77	48.33	0.106	24.90	52.69	0.198	25.99	52.47	0.225
Qwen-3-VL	38.50	61.80	0.310	14.20	26.50	0.085	22.40	34.60	0.155	24.30	36.80	0.190

Table 15: Performance analysis across reasoning types. EM: Exact Match, SS: Substring Match, F1: F1 Score

D Human Evaluation and Error Analysis

We conducted human evaluation with two annotators on 50 tables (11% of total dataset) and 393 questions (12% of total questions), where participants answered based solely on their knowledge without external image search capabilities. Results show humans consistently outperform all baseline models, achieving exact match scores in the 75-85% range and substring scores above 80%, demonstrating strong human performance while still revealing the inherent complexity of multimodal table reasoning tasks.

For the Entity Replaced baseline (upper bound), humans would perform exceptionally well since all required information appears in text format. For the Missing Image baseline (lower bound), models might outperform humans as they can compensate for missing information using pre-training knowledge, unlike human participants. Therefore, for fair comparison, we conduct in-depth analysis comparing the best-performing model across Table as Image and Interleaved baselines with human performance.

D.1 Types of Errors in Image Interpretation

We performed error evaluation and manually classified each error into one of the following categories:

- **Entity Disambiguation Issues** – These errors occur when an image is incorrectly identified, leading to misinterpretation of its content. Misidentification can lead to entirely incorrect conclusions about the image.
- **Entity Identification Issues** – This error refers to the complete failure to recognize or identify the image.
- **Reasoning Errors** – This category includes mistakes where the image is correctly identified, but the logic used to answer the question is flawed. Such errors typically involve incorrect inferences, faulty assumptions, or logical inconsistencies.
- **Identification of Visual Attributes** – Errors in this category involve the failure to recognize key visual components of an image. This could mean missing out on crucial details such as shapes, colors, textures, or patterns.
- **Structural Errors** – These errors specifically pertain to the misinterpretation of tabular or structured data within an image. Failure to correctly identify rows, columns, and hierarchical relationships can result in inaccurate conclusions when analyzing tables, charts, or diagrams.
- **Mathematical Errors** – These occur when an individual or AI system miscalculates numerical information within the table. Errors can include

Model	SE		ME		SN		MN		IL		MT	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
Missing Image Baseline												
Gemini 1.5 Flash	18.9	0.071	21.1	0.081	13.3	0.037	20.4	0.084	15.4	0.028	22.1	0.112
Gemini 2.0 Flash	19.0	0.084	16.9	0.094	15.8	0.037	24.8	0.090	2.6	0.014	25.9	0.067
GPT-4o Mini	32.0	0.230	32.0	0.201	27.4	0.188	33.0	0.224	10.6	0.073	37.3	0.240
Llama 3-8B	28.1	0.188	29.7	0.182	21.3	0.137	29.1	0.145	0.0	0.000	25.4	0.160
Mixtral	35.3	0.273	33.3	0.247	28.4	0.209	36.7	0.233	3.4	0.025	38.0	0.316
Entity Replaced Baseline												
Gemini 1.5 Flash	37.8	0.114	57.4	0.385	7.2	0.615	25.7	0.193	8.0	0.300	36.7	0.308
Gemini 2.0 Flash	41.7	0.211	59.0	0.315	15.8	0.558	30.8	0.248	11.0	0.167	16.7	0.257
GPT-4o Mini	48.1	0.290	66.9	0.481	22.8	0.598	36.8	0.334	13.4	0.227	24.1	0.368
Llama 3-8B	39.8	0.210	57.5	0.412	16.2	0.504	31.7	0.308	13.8	0.106	14.2	0.317
Mixtral	41.3	0.293	56.6	0.446	29.7	0.593	35.0	0.348	14.8	0.205	27.7	0.350
Image Captioning Baseline												
Gemini 1.5 Flash	28.3	0.217	22.7	0.155	25.3	0.188	12.9	0.051	5.8	0.058	14.7	0.056
Gemini 2.0 Flash	33.3	0.243	34.5	0.230	25.8	0.183	23.2	0.109	4.0	0.029	29.6	0.131
Table As Image Baseline												
Gemini 1.5 Flash	38.0	0.223	41.6	0.274	21.6	0.078	46.6	0.255	23.2	0.113	34.3	0.199
Gemini 2.0 Flash	30.3	0.147	35.6	0.168	19.7	0.072	38.6	0.174	26.9	0.119	32.5	0.149
GPT-4o Mini	47.0	0.355	46.5	0.305	33.4	0.228	50.1	0.318	26.7	0.214	42.5	0.304
Intern-VL-2.5	17.7	0.187	18.9	0.189	12.1	0.125	27.6	0.292	8.6	0.080	23.0	0.227
Mantis-8B-Idefics2	22.6	0.140	23.8	0.106	10.7	0.040	24.6	0.088	25.6	0.122	22.7	0.083
Phi-3.5	18.2	0.089	21.2	0.111	12.5	0.038	17.6	0.023	0.1	0.002	14.0	0.056
Qwen-2.5-VL	29.7	0.160	32.1	0.190	15.3	0.059	29.8	0.161	14.1	0.067	29.2	0.181
Qwen-3-VL	38.5	0.265	39.8	0.250	24.0	0.145	40.2	0.245	21.0	0.145	36.5	0.245
Table_LLaVA	12.9	0.078	14.0	0.050	10.8	0.030	21.9	0.065	0.0	0.000	16.6	0.050
Interleaved Baseline												
Gemini 1.5 Flash	29.2	0.205	29.0	0.177	25.2	0.175	25.2	0.120	4.5	0.034	26.6	0.131
Gemini 2.0 Flash	33.2	0.223	34.6	0.210	27.9	0.195	29.4	0.156	26.6	0.131	27.3	0.103
GPT-4o Mini	48.1	0.372	47.7	0.301	30.9	0.223	40.0	0.252	21.6	0.127	24.4	0.184
Mantis-8B-Idefics2	26.0	0.175	29.2	0.144	15.6	0.064	15.8	0.061	27.0	0.139	21.7	0.066
Phi-3.5	21.4	0.136	27.7	0.144	13.7	0.060	28.8	0.131	13.7	0.071	27.0	0.185
Qwen-2.5-VL	29.4	0.261	21.5	0.177	23.5	0.185	22.0	0.158	5.7	0.060	20.2	0.165
Qwen-3-VL	39.0	0.320	35.5	0.240	27.0	0.205	31.0	0.210	19.5	0.105	23.8	0.175

Table 16: Performance across answer types. SE: Single Entity, ME: Multiple Entity, SN: Single Number, MN: Multiple Number, IL: Image Location, MT: Multiple Types.

Baseline	Model	Answer Mention Questions			Explicit Question			Implicit Question			Visual-Based Question		
		Exact	Substring	F1	Exact	Substring	F1	Exact	Substring	F1	Exact	Substring	F1
Table as an Image Baseline	GPT-4o-mini	34.8	38.7	0.32	44.8	46.9	0.34	40.7	42.3	0.32	41.0	45.2	0.36
Interleaved Baseline	GPT-4o-mini	41.6	45.9	0.35	46.4	47.8	0.37	31.4	33.0	0.26	32.2	34.7	0.27
Interleaved Baseline	Human answers	78.4	82.1	0.76	84.2	87.3	0.81	75.8	80.6	0.73	79.9	83.7	0.78

Table 17: Best performing baselines on sample set evaluation.

incorrect counting, misreading of numerical values, or computational mistakes when deriving conclusions from visual data that involves numbers.

- **Partial Answers** – This type of error involves providing an incomplete response where some crucial details are missing. The responder may identify and interpret part of the image correctly but fail to provide a comprehensive answer.
- **Extra Information or Hallucination** – This occurs when irrelevant or incorrect details are introduced into the response. The additional information may not be present in the table itself but might be inferred incorrectly based on prior pre trained knowledge.

Error Hierarchy and Root Cause Attribution. While a single response can exhibit multiple types of errors, we adopt a hierarchical approach to identify the *root cause* of failure. This hierarchy ensures that derivative errors (e.g., reasoning or partial answers resulting from earlier misinterpretations) are not double-counted.

During analysis, if an instance exhibits multiple error types, we attribute it only to the **root cause**. For example, if a model fails to recognize a visual attribute and consequently provides an incomplete answer, the error is classified under *Identification of Visual Attributes*, not as a reasoning or partial answer error.

The analysis highlights key patterns in human performance, with Entity Identification Issues being the most common, suggesting difficulties in recognizing entities, particularly without domain knowledge. Partial Answers reveal a tendency to overlook essential details, while Entity Disambiguation Issues indicate occasional struggles in distinguishing between similar entities. These three challenges emphasize the importance of domain knowledge in accurately solving tasks and identifying the correct entity.

Visual Attribute Identification Errors suggest that even humans can miss fine details in images. However, Reasoning Errors are relatively low, in-

Type of Error	Human (%)	Interleaved (%)		Table as Image (%)	
		GPT-4o-mini	Qwen2.5-VL	GPT-4o-mini	Qwen2.5-VL
Entity Identification Issues	74.44	13.23	8.42	18.67	16.74
Entity Disambiguation Issues	5.21	11.00	7.85	8.00	6.27
Identification of Visual Attributes	11.32	17.08	12.78	16.00	9.38
Partial Answer	6.02	12.08	13.32	14.00	15.87
Reasoning Errors	3.01	14.38	16.11	15.33	17.98
Structural Errors	0.00	11.15	13.46	7.33	12.91
Mathematical Errors	0.00	10.31	11.09	8.67	9.59
Extra Info / Hallucination / Pretrained Knowledge	0.00	10.77	16.97	12.00	11.26
Total	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**

Table 18: Error type distribution across evaluation methods.

dicating that humans generally follow logical processes. Notably, there are no Mathematical Errors, Structural Errors, or Hallucinations, demonstrating strong numerical reasoning and structured data interpretation.

Overall, humans excel in logical analysis and structured comprehension but face challenges in entity recognition, visual attention to detail, and providing complete answers.

D.2 Performance and Analysis of Errors in Interleaved baseline GPT 4o-mini

GPT-4o-mini exhibits significant errors across multiple categories, with each exceeding the 10% threshold. Reasoning Errors are the most prominent, as the model struggles with multi-step deductions and logical coherence - an ongoing challenge for MLLMs. Visual Attribute Identification is another major limitation, with failures in extracting key image features, underscoring the need for better vision encoders and fine-tuning.

Entity Identification and Disambiguation errors occur at similar rates, as the model misidentifies or fails to recognize entities, likely due to insufficient training data and over-reliance on context. Structural Errors show difficulty in interpreting complex tabular data, including hierarchical structures and nested tables, while Hallucination Errors highlight the model's tendency to generate irrelevant information. Mathematical Errors reveal persistent struggles with quantitative reasoning.

These issues demonstrate that while GPT-4o-mini has some multimodal reasoning capability, it lacks depth in entity recognition, visual interpretation, and structured data comprehension.

D.3 Performance and Analysis of Errors in Table as an Image baseline GPT 4o-mini

GPT4o-mini makes many errors across all categories. From table 18, in table as image baseline, we notice that GPT4o-mini had most errors in Entity Identification. This is likely due to insufficient training data to much reliance on context.

It also struggles with Identification of visual attributes and reasoning errors in similar range as Interleaved baseline. We see similar error rates among Mathematical, Hallucinations, Partial and Entity Disambiguation error types showing that GPT4o-mini struggles with these issues irrespective of input table format.

Interestingly, We observe that structural Errors decreased by ~4% in table as image baseline. This may be due to the fact that looking at entire table at once through image helped understand the structure better. So, passing tables as image does help GPT4o-mini to understand the structure of tables better.

From above both analysis we can say GPT4o-mini has scope of improvement in identifying visual attributes, on entity identification and reasoning errors.

D.4 Performance and Analysis of Errors in Interleaved baseline Qwen2.5-VL

The error distribution in Qwen2.5-VL suggests a model that demonstrates broad multimodal competence but lacks consistent grounding in visual inputs. The high incidence of hallucination-related errors (16.97%) implies that the model frequently relies on pretrained textual priors rather than accurately integrating perceptual evidence. This points to limitations in the vision-language alignment mechanisms, where generated responses may reflect plausible but unsubstantiated inferences.

Reasoning errors (16.11%) and structural errors (13.46%) indicate significant challenges in multi-step logical processing and the interpretation of complex visual structures, such as nested tables or hierarchical layouts. These errors suggest that while Qwen2.5-VL is able to extract surface-level

patterns, it struggles with deeper, context-sensitive reasoning tasks - an area where multimodal systems continue to underperform.

Although errors in entity identification (8.42%) and disambiguation (7.85%) are comparatively lower, they reflect persistent difficulties in maintaining contextual precision, particularly in visually dense or ambiguous scenarios. Overall, the model's performance underscores the need for improved grounding strategies, enhanced visual encoding, and tighter integration of reasoning capabilities to support more reliable and context-aware multimodal understanding.

D.5 Performance and Analysis of Errors in Table as an Image baseline Qwen2.5-VL

Qwen2.5-VL's performance on table-as-image tasks highlights persistent challenges in structured data interpretation. The most prominent error category is Reasoning Errors (17.98%), suggesting the model struggles to perform logical operations over tabular data, particularly when spatial layout or implicit relationships between rows and columns are involved. This reflects a broader difficulty in translating visual table structures into coherent semantic representations.

Partial Answers (15.87%) and Entity Identification Issues (16.74%) further indicate that Qwen2.5-VL often fails to capture complete and precise information from table-based inputs. These errors likely stem from limitations in visual attention mechanisms or token alignment between image features and textual output. Similarly, the model's relatively high rate of Visual Attribute Identification errors (9.38%) and Structural Errors (12.91%) reveals an incomplete understanding of layout hierarchies, cell groupings, and formatting cues critical for accurate parsing.

While hallucination (11.26%) and mathematical errors (9.59%) are somewhat less frequent, they still suggest inconsistencies in quantitative reasoning and information grounding. Collectively, these findings indicate that despite some robustness in basic vision-language alignment, Qwen2.5-VL requires substantial improvements in handling visually structured data, particularly for tasks requiring deep reasoning and layout-aware comprehension.

E Prompts

This section contains the exact prompts used in our experiments.

Missing Image Baseline Prompt

You will be provided a table in a pipe-separated table where all the entities have been removed. Your task is to:

Step 1: UNDERSTAND THE TABLE CONTEXT - Carefully analyze the table structure and identify its purpose and what it mentions.

Step 2: FILL IN THE GAPS - Use the table context and your real-world knowledge to deduce the missing entities logically.

Step 3: ANALYZE THE QUESTIONS - Read all the questions provided and explore ***ALL TYPES OF REASONING*** to find answers, including but not limited to Numerical reasoning(relationships, totals, and comparisons), Visual reasoning (Colors, shapes, or patterns), Contextual reasoning, (Real-world connections or logic), etc.

Step 4: PROVIDE ANSWERS IN FORMAT - Ensure that all answers adhere strictly to the FORMAT specified. Avoid deviating from this format or including unnecessary explanations.
{Answer Formatting Guidelines}
ALWAYS PROVIDE YOUR ANSWERS IN THIS FORMAT.

If you are unable to answer it, simple answer *UNKNOWN*.

I will provide one example to show you:
{One Shot Example}
Now I will provide you with the table and question.
{TABLE}
{QUESTION}

Based on the examples that I have provided and the steps I mentioned above, answer the question.

Entity Replaced Baseline Prompt

You will be provided a pipe-separated table format that contains some entities. Your task is to:

Step 1: UNDERSTAND THE TABLE CONTEXT - Carefully analyze the table structure and identify its purpose and what it mentions.

Step 2: ANALYZE THE QUESTIONS - Read all the questions provided and explore **ALL TYPES OF REASONING** to find answers, including but not limited to Numerical reasoning (relationships, totals, and comparisons), Visual reasoning (colors, shapes, or patterns), Contextual reasoning (real-world connections or logic), etc.

Step 3: PROVIDE ANSWERS IN FORMAT - Ensure that all answers adhere strictly to the FORMAT specified. Avoid deviating from this format or including unnecessary explanations.
{Answer Formatting Guidelines}
ALWAYS PROVIDE YOUR ANSWERS IN THIS FORMAT.

IMPORTANT ALL answers are there in the table. ANALYZE the question and table properly.

I will provide one example to show you:
{One Shot Example}
Now I will provide you with the table and question.
{TABLE}
{QUESTION}

Based on the examples that I have provided and the steps I mentioned above, answer the question.

Image Captioning Baseline Prompt

You will be provided a table in a pipe-separated table with images included. Your task is to:

Step 1: UNDERSTAND THE TABLE CONTEXT - Carefully analyze the table structure and identify its purpose and what it mentions.

Step 2: CAPTION EVERY IMAGE - Based on the image, provide a caption for that image. Your job is to reason, predict, and replace image entity tags and provide visual descriptions.

Step 3: CREATE A TABLE - Based on the image captions, create a pipe-separated table where the image placeholders or cells have been replaced with their captions.

Step 4: ANALYZE THE QUESTIONS - Read all the questions provided and explore **ALL TYPES OF REASONING** to find answers.

{TABLE WITH CAPTIONS}

Step 5: PROVIDE ANSWERS IN FORMAT Ensure that all answers adhere strictly to the FORMAT specified.
{Answer Formatting Guidelines}
ALWAYS PROVIDE YOUR ANSWERS IN THIS FORMAT.

IMPORTANT ALL answers are there in the table. ANALYZE the question and table properly.

Now I will provide you with the question.
{QUESTION}

Based on the steps I mentioned above, answer the question.

Table as an Image Baseline Prompt

You will be provided an image of a table. Your task is to:

Step 1: UNDERSTAND THE IMAGE CONTEXT - Carefully analyze the image content and understand the tabular structure and all text and visual aspects inside the image.

Step 2: ANALYZE THE QUESTIONS - Read all the questions provided and explore **ALL TYPES OF REASONING** to find answers.

Step 3: PROVIDE ANSWERS IN FORMAT Ensure that all answers adhere strictly to the FORMAT specified. Avoid deviating from this format or including unnecessary explanations.
{Answer Formatting Guidelines}
ALWAYS PROVIDE YOUR ANSWERS IN THIS FORMAT.
IMPORTANT ALL answers are there in the image.

Now I will provide you with the image of the table.
{IMAGE OF TABLE}
For this image, you will answer the following question.
{QUESTION}
Based on the steps I mentioned above, answer the question.

Interleaved Baseline Prompt

You will be provided a pipe-seperated table where some cells are images. Your task is to:

Step 1: UNDERSTAND THE TABLE CONTEXT - Carefully analyze the table structure and understand the intricate relationship between image and text.

Step 2: ANALYZE THE QUESTIONS - Read all the questions provided and explore **ALL TYPES OF REASONING** to find answers.

Step 3: PROVIDE ANSWERS IN FORMAT Ensure that all answers adhere strictly to the FORMAT specified. Avoid deviating from this format or including unnecessary explanations.
{Answer Formatting Guidelines}
ALWAYS PROVIDE YOUR ANSWERS IN THIS FORMAT.

IMPORTANT ALL answers are there in the image.

Now I will provide you with the table.
{INTERLEAVED TABLE}
For this table, answer the following question.
{QUESTION}
Based on the steps I mentioned above, answer the question.

Answer Formatting Guidelines The following answer formatting guidelines were provided along with every prompt to eliminate inconsistencies and ensure a uniform response structure across all task types. The model is expected to provide only the final answer, formatted as described below:

- **Single Entity:** Return a single *string* representing one entity such as a name, country, company, object, or similar. The answer should be concise and written in one line without extra text.
  - Example (Name): `Elon Musk`
  - Example (Country): `China`
  - Example (Company): `Google`
  - Example (Color): `Red`
- **Single Number:** - If the answer is a whole number, write it **without decimals**. - If it

has decimals, round to **two decimal places**. - If the last digit after rounding is 0 (e.g., 23.40), remove the trailing zero ($\rightarrow$ **23.4**). Units should only be included if explicitly mentioned in the question.
  - Example (Whole Number): `45`
  - Example (Decimal): `12.36`
  - Example (Trimmed Decimal): `23.4`
- **Multiple Entities:** Provide a **list of strings**, each following the same rules as the Single Entity format. Use comma-separated values enclosed in square brackets.
  - Example: `["Apple", "Microsoft", "Google"]`
- **Multiple Numbers:** Provide a **list of numbers**, each following the Single Number formatting rule. Use comma-separated values enclosed in square brackets.
  - Example: `[23, 45.67, 89.4]`
- **Image Locations:** When the answer involves identifying a location within a visual or tabular structure, specify it using the following format: `row_num_col_num`.
  - Example: `row_2_col_3`