Knowledge-Aware Reasoning over Multimodal Semi-structured Tables

Suyash Vardhan Mathur[1], Jainit Bafna[1*], Kunal Kartik[2*], Harshita Khandelwal[3*]
Manish Shrivastava[1], Vivek Gupta[4†], Mohit Bansal[5], Dan Roth[6]
[1]IIIT Hyderabad, [2]IIT Guwahati, [3]UCLA, [4]ASU, [5]UNC Chapel Hill, [6]UPenn
{suyash.mathur, jainit.bafna}@research.iiit.ac.in; harshitaskh@g.ucla.edu; kunalkartik02@gmail.com;
m.shrivastava@iiit.ac.in; vgupt140@asu.edu; mbansal@cs.unc.edu; danroth@seas.upenn.edu

Abstract

Existing datasets for tabular question answering typically focus exclusively on text within cells. However, real-world data is inherently multimodal, often blending images such as symbols, faces, icons, patterns, and charts with textual content in tables. With the evolution of AI models capable of multimodal reasoning, it is pertinent to assess their efficacy in handling such structured data. This study investigates whether current AI models can perform knowledge-aware reasoning on multimodal structured data. We explore their ability to reason on tables that integrate both images and text, introducing MMTabQA, a new dataset designed for this purpose. Our experiments highlight substantial challenges for current AI models in effectively integrating and interpreting multiple text and image inputs, understanding visual context, and comparing visual content across images. These findings establish our dataset as a robust benchmark for advancing AI's comprehension and capabilities in analyzing multimodal structured data.

1 Introduction

Tables are crucial for efficiently summarizing and conveying information across various fields. In real-world applications, they often include images representing entities, such as team logos in sports scoreboards and product features in E-commerce tables (Fig. 1). In medicine, tables may display visual symptoms for comparing diseases, while educational tables might include molecular diagrams or images of plant species. Wikipedia tables frequently incorporate images, such as team logos in sports articles or comparative tables for scientists, Nobel laureates, and ship classes. Political party tables often feature election symbols and charts illustrating seat wins. This integration of images enriches the data's depth and informativeness.

Q1 How to unlock the phone which has a dual horizontal camera?
A1 Fingerprint Scanner
Q2 Which phone combine three camera lens with latest processor?
A2 iPhone 11
Q3 Which phone comes with the fewest color options?
A3 iPhone 8 or 8+

Figure 1: Multimodal Table comparing iPhone features

Understanding and interpreting these multimodal tables is crucial across various domains. In healthcare, they help doctors compare disease symptoms for accurate diagnosis and treatment planning. In education, students use visual aids in tables to better understand complex concepts. In E-commerce, consumers rely on product comparison tables to make informed purchasing decisions.

Advances in modeling techniques, including table pre-training and targeted fine-tuning, have greatly improved reasoning capabilities for semi-structured tables (Müller et al., 2021; Aly et al., 2021). Furthermore, large language models (LLMs) have shown remarkable performance across diverse domains, achieving state-of-the-art performance on various tabular reasoning tasks (Chen et al., 2021; Wang et al., 2021; Lu et al., 2023). Despite extensive exploration of inference and reasoning over tables most prior works (Jin et al., 2022) has primarily focus on text-only tables.

*equal contribution. †corresponding author (work done while at UPenn)

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14054–14073
November 12-16, 2024

Thus, developing advanced AI models to process multimodal tables is essential. These models must integrate textual and visual data for comprehensive analysis across fields. Improving these models can enhance the accuracy and efficiency of tasks in healthcare diagnostics, education, and consumer decision-making, thereby enriching data presentation and enhancing its informativeness and utility.

Reasoning with multimodal tables poses significant challenges. Table reasoning, as illustrated in Fig. 1, necessitates entity disambiguation within the table context. For instance, disambiguating the A13 square as representing the latest A13 processor chip is essential for answering Q2. Additionally, visual reasoning over individual images in the table is crucial, such as using the phone images to determine the camera alignment for answering Q1. Similarly for Q3, understanding phone colors through their visual representation and counting them could be challenging. To answer these questions well, model must understand the images in the table and its relation to other cells (images and text), which involves complex reasoning such as visual analysis, numerical interpretation, temporal sequencing, and entity relationship identification.

However, the lack of datasets to evaluate Vision-Language models on multimodal table reasoning has left this area largely unexplored. Therefore, this paper aims to investigate the research question: *Can current Vision-Language models handle complex reasoning in multimodal tables?* To effectively tackle this challenge, we introduce a new task called *knowledge-aware reasoning over multimodal semi-structured tables*. Due to the time-consuming and costly process of curating a new human-annotated dataset on multimodal tables, we repurpose existing Wikipedia datasets into a multimodal format. Our framework replaces recognizable entities in textual Wikipedia tables with their representative images, creating the MultiModal TABle Question Answering (MMTABQA) dataset repurposed using four Wikipedia tables-based question-answering datasets.

In MMTABQA, we categorize questions into explicit (mentioning an image-replaced entity explicitly), answer-mention (referencing an image-replaced entity in the answer), and implicit (involving image-replaced entities in intermediate reasoning). We also generate synthetic visual questions by enhancing explicit questions with visual attributes of the mentioned entity and validate them through human evaluation. Evaluating various state-of-the-art closed and open-source LLMs and VLMs using diverse modeling approaches on MMTABQA reveals challenges in entity disambiguation, understanding table structures, and performing visual reasoning. We aim for our dataset to be a robust benchmark for evaluating Vision-Language Models (VLMs) on complex multimodal tabular reasoning. We summarize our contributions as below:

- We propose Knowledge-Aware Reasoning over Multimodal Semi-structured Tables and present a framework to repurpose Wikipedia textual-table datasets for multimodal tasks.
- Using this framework, we create the MMTABQA dataset for studying knowledge-aware multimodal reasoning over tables and evaluate various Vision-Language Models (VLMs) using diverse techniques.
- Our analysis shows that current VLMs face challenges in performing reasoning on MMTABQA. They struggle with erroneous entity linking, visual understanding difficulties, and table structure comprehension.

The dataset, along with associated scripts, are available at `https://mmtabqa.github.io/`.

2 Multimodal Tabular Reasoning

Today's VLMs face multiple challenges when reasoning with multimodal tabular question-answering datasets. Below, we describe these challenges in detail:

2.1 Tabular Multimodal Structure

Table reasoning is inherently challenging as it needs to rightly interpret semi-structured data, understand complex entity relationships, and integrate diverse contexts (Fang et al., 2024). This difficulty is compounded when processing even a single image with text as an additional modality (de Faria et al., 2023). Multimodal table reasoning involves multiple images, while current VLMs are optimized to reason over a single image. Unlike a separated context, these images are semi-structured within the table context. In Fig. 1, phone images in the table correspond to the named phones in the header, and processor images must be linked to their respective phone columns. Encoding such tables as interleaved multimodal inputs is particularly challenging for VLMs (Tian et al., 2024). Thus, exploring various approaches, such as captioning images individually to create a text-only table or representing the entire table as an image, can be further explored.

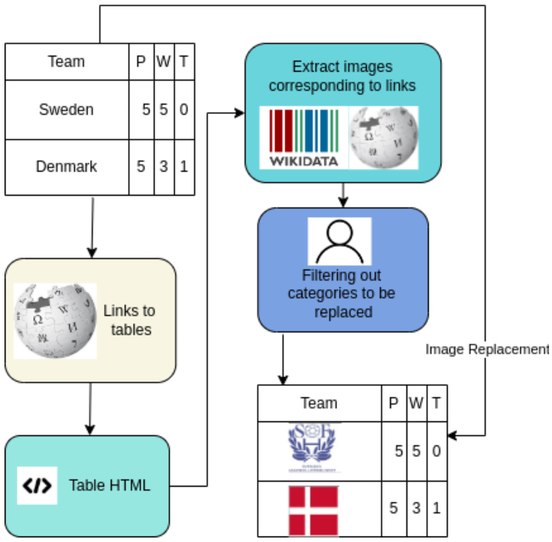


Figure 2: Dataset Creation Pipeline

2.2 Complex Multimodal Reasoning

Beyond understanding the multimodal table, the model must perform complex reasoning to answer questions. For instance, in Fig. 1, answering *"Which iPhones have the A13 processor?"* requires the model to identify the A13 chip and link the image of a fingerprint to a fingerprint scanner, generating the corresponding text. This task is more challenging than intermediate reasoning, especially when questions reference image-replaced entities. The table's context, such as comparing iPhone features, is crucial for accurate entity disambiguation. Additionally, answering questions often requires comparing visual attributes across images, demanding a visual understanding beyond simple entity disambiguation. For example, to answer Q1, the model must compare camera placement across all phone images. Moreover, some questions involve reasoning over multiple images, further complicating the task. Answering Q1 requires comparing camera alignment over several phone images, and other questions may require complex reasoning, such as temporal, numerical, and entity reasoning, to derive the correct answer.

3 MMTabQA Dataset

3.1 Original Tabular Dataset

To diversify the dataset for knowledge-based question answering over multimodal tables, we adapted four existing Wikipedia-based datasets each featuring a variety of real-world entities. Specifically, we adapted the following datasets:

- WikiSQL Dataset (Zhong et al., 2017) to benchmark model capabilities in parsing entities accurately and answering basic SQL-based questions.
- WikiTableQuestions dataset (Pasupat and Liang, 2015) to include questions which require more complex reasoning.
- FeTaQA dataset (Nan et al., 2022) to include long-form answer based questions which involve multiple row/column reasoning.
- HybridQA dataset (Chen et al., 2020) which includes extra contextual passages beyond the tables, requiring hybrid complex reasoning.

3.2 MMTabQA Creation

MMTabQA Table Generation

To convert tables from textual to multimodal form, we link textual entity mentions to corresponding images. Wikipedia datasets enable easy access to entity images from Wikipedia Infoboxes or Wikidata entries using Wikipedia page links.

For each dataset, the raw HTML of the table's page is obtained using the corresponding page revisions and the *Jaccard Coefficient* is used to locate the table on the page. The Wikipedia links from the raw HTML table are extracted with their corresponding images, prioritizing Wikipedia Infobox images and using a defined priority order for Wikidata images (Lerner et al., 2022). We filter entities for image replacement based on the WikiData P31 "instance of" property and corresponding Wikipedia pageviews to find popular and recognizable images. Over 1,500 unique "instance of" values were annotated for pageview-based filtering across the datasets, while limiting certain values to only use seals, coat of arms, or logo images from the infobox.

Finally we replace linked text in tables with corresponding images from their Wikipedia URLs. In WikiTableQuestions, linked text is directly replaced with scraped images using the original HTMLs provided. For the WikiSQL and FeTaQA datasets, all $(link, text)$ pairs are extracted and filtered from the HTML, and the text in the original table is replaced. In HybridQA, we leverage provided Wikipedia links corresponding to different cells to enhance table quality when replacing text. Additionally, coreference resolution is used to find all mentions of image-replaced entities in their passages in HybridQA, and they are replaced with tags to prevent entity name leaking.

MMTabQA **Question Filtering & Creation**

We filter out the questions corresponding to the tables of the following types:

Explicit Questions which mention an entity that is replaced by an image in the table.

Answer-Mention Questions whose answer contains an entity that is replaced by an image in the table, but the question does not.

Implicit Questions where an image-replaced entity is involved in intermediate reasoning but not mentioned in the answer or the question.

We use simple string matching for filtering out the explicit and answer-mention questions and use evidence cells for filtering implicit questions. Finally, we exclude tables lacking a column with at least 30% images and cap each column at 75% images to maintain a balance between images and text cells, prioritizing retaining evidence cells.

We introduce **visual questions** in our dataset, involving visual aspects of entity images from the table by recasting the explicit mention questions. Visual question creation is limited to specific image categories: landscape collage, logo, seal, flag, coat of arms, and poster. For each image and other images of the same category in the table, VLM (Gemini 1.0 (Team et al., 2023)) is prompted for category-specific visual attributes. LLM (Gemini) is then prompted to provide a set of unique attributes for the explicit entity's image and to replace its mention in the question with these attributes.

3.3 MMTabQA Validation

Tables Validation Since we recast existing tables, we first need to verify whether the entity replacements are correct. We sample 250 tables from each data source (total 1000 tables) and have 3 annotators score all unique $(image, original\_text)$ pair per table, verifying the correctness of the image replacements. Label 0 indicates that the image used for the entity is incorrect; Label 1 indicates the image represents the entity but is ambiguous for a human to identify; Label 2 indicates the image clearly represents the entity We report the annotation results in Table 1.

Questions Validation Explicit, answer-mention, and implicit questions are repurposed from existing tabular QA datasets and do not require additional validation, as the tables themselves are already validated. However, we need to validate the synthetically created visual questions. Three annotators score 500 recast questions as 0 or 1, where 0 indicates that the recast question is incorrect (wrong attribute hallucinated, not uniquely identifiable with table attributes) while 1 indicates appropriate question. We use inter-annotator agreement and obtain 15.6% questions annotated as 0 while 84.4% examples are annotated as 1.

Data Source	**No Agree.**	**0**	**1**	**2**
FeTaQA	0.28(14)	0.00(0)	0.08(4)	99.64(5030)
HybridQA	0.46 (28)	0.26 (16)	2.20 (134)	97.08 (5910)
WikiSQL	0.00 (0)	0.10 (6)	0.04 (2)	99.86 (5688)
WikiTable-Questions	0.43 (40)	0.26 (24)	0.13 (12)	99.19 (9282)

Table 1: MMTabQA Agreement Statistics: The number in brackets represent absolute number.

3.4 MMTabQA Statistics

As described, we create the MMTabQA dataset with 69,740 questions over 25,026 tables. Major statistics are in Table 2. Additional statistics about the data distribution and content are presented in Appendix A.

Data Source	**No. of Questions**	**No. of Tables**	**Avg. Img per Table**
WikiSQL	21,472	9,784	13.68
WikiTable-Questions	10,052	1,259	17.67
FeTaQA	7,476	5,898	10.43
HybridQA	30,470	8,085	14.64
Overall	69,740	25,026	14.10

Table 2: MMTabQA Statistics

4 Modelling Strategies

To benchmark the model performance on our MMTabQA dataset, we define four baselines:

4.1 Partial Input Baseline

In this baseline, images are excluded, providing only the table with replaced image tags alongside the question to the model. These image tags act as placeholders indicating where images would be in the text format. The model makes guesses about which entities correspond to the image/entity tags for QA. This baseline serves as a lower bound, as models with direct access to images are expected to perform better.

4.2 Image-captioning Baseline

Here, the table is converted to a text-only format for reasoning using Language Models (LLMs), using captions generated by VLMs instead of image tags.

- **Entity Prediction:** Initially, we predict entities for each image occurrence using infobox-style tables. These tables are created with individual rows of the table, where cell corresponding to same column as entity of interest is text-only. VLMs then use this context to predict the original text associated with each image along with a brief visual description of the image.

- **Question Answering:** After preparing table T, question Q, and predicted entities E with their visual descriptions V, we prompt LLM to generate the answer to Q using T while considering V and E, explicitly describing possible inaccuracy of V and E.

Captioning individual images within tables is highly resource-intensive, especially since tables typically contain 10-16 images each. Despite its computational expense, this baseline is valuable for converting the task into text-only format, enabling the use of a larger LLM to handle complex reasoning and integrate visual and textual data effectively.

4.3 Table-Image Baseline

Here, we create an image of the table that includes all embedded entity images. This multimodal input, consisting of the table image and textual question, is directly inputted into the model.

4.4 Interleaved Image-text Baseline

This baseline fully integrates both visual and textual modalities, providing a comprehensive representation. Unlike the first two baselines, which compromise the visual input, and the Table-Image Baseline, which makes textual reasoning challenging, this model achieves optimal representation by combining both modalities effectively. To encourage prompt understanding, LLMs are employed to perform row pruning on tables before evaluating open-source models.

4.5 Oracle-Entity Replaced Baseline

We also evaluate oracle entity-replaced textual tables, which are the original textual tables from which the multimodal tables were derived. We do not report these numbers for visual questions because mere entity replacement is inadequate for addressing such questions. This baseline sets an upper bound for explicit, answer-mention, and implicit questions in our dataset, representing the model's performance when entity disambiguation from the image is perfectly executed for tabular reasoning.

5 Experiments

Models used We employ a combination of open-source and closed-source models to benchmark performance on our dataset. Specifically, Google's closed-source Gemini 1.5 Flash & Open AI's GPT-4o (Achiam et al., 2023) are utilized for the majority of modeling approaches. For the textual Oracle-Entity Replacement baseline and the Partial Input baseline, open-source textual LLMs, namely LLaMa-3 70B (Touvron et al., 2023) and Mixtral 8x7B (Jiang et al., 2024a), are employed. For the table-as-image and interleaved baselines, open-sourced Qwen-VL Chat (Bai et al., 2023) is used. Additionally, CogAgent-VQA (Hong et al., 2023) and Intern-VLM-xComposer-4khd (Chen et al., 2024) are benchmarked for the table-image baseline, and the Idefics-Mantis model (Jiang et al., 2024b), specifically trained for handling multiple interleaved images, is used for the interleaved baseline. Due to resource constraints, only Gemini 1.5 Flash is run on the Table Captioning baseline.

Our prompting methodology combines few-shot learning (Brown et al., 2020) and Chain of Thought (COT) prompting (Wei et al., 2023). COT aids the model in understanding the reasoning process behind a question, while few-shot examples guide the expected answer format. Depending on the model and available resources, we use either 4 or 8 examples for prompting. Sample prompts can be found in Appendix G.

Evaluation Metrics We used different metrics for the two types of answers in our tasks. For single-word or phrase answers, like those in WikiTableQuestions and WikiSQL, we used Substring Match, which checks for the correct answer within the predicted text. For long-form or sentence answers, like those in FeTaQA, we used ROUGE-L (Lin, 2004), which evaluates the longest common subsequence between predicted and reference texts. Detailed evaluations for each data source and question type, with multiple metrics, are in the appendix.

Evaluation Benchmark We sample out 20% questions per dataest per question type, sampling at least 500 questions and maximum of 700 questions for our test set.

6 Result and Analysis

We observe that our models' performance varies significantly across different datasets, methods, and models. We provide a detailed analysis of these variations below:

Dataset	WikiTableQuestions				WikiSQL				FetaQA			
Model	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ	EQ	AQ	IQ	VQ
Partial Input Baseline												
Gemini-1.5 Flash	40.99	27.38	48.95	31.4	39.14	28.71	62.22	28	0.51	0.44	0.44	0.47
GPT-4o	57.45	38.02	70.83	42.40	52.57	43.86	72.38	39.00	0.51	0.46	0.42	0.44
Llama-3 70B	41.13	26.48	43.75	31.8	41.117	30.75	61.27	30.6	0.52	0.46	0.45	0.48
Mixtral 8x7B	26.56	9.90	30.26	20.2	23.42	17.71	28.88	19.2	0.44	0.39	0.38	0.39
Oracle-Entity Replaced Baseline												
Gemini-1.5 Flash	74.89	78.19	54.86	-	82.28	81.86	77.46	-	0.56	0.50	0.41	-
GPT-4o	87.80	84.86	84.55	-	85.57	82.71	79.05	39.00	0.53	0.48	0.43	-
Llama-3 70B	75.74	75.31	58.85	-	78.28	78.57	68.25	-	0.49	0.46	0.41	-
Mixtral 8x7B	54.89	53.87	40.69	-	59.28	69.28	33.96	-	0.44	0.41	0.33	-
Image-Captioning Baseline												
Gemini-1.5 Flash	52.34	42.16	51.39	42.2	50.42	40.85	67.30	46.6	0.57	0.46	0.42	0.43
Table-as-an-Image Baseline												
Gemini-1.5 Flash	44.22	25.65	41.01	37.8	47.08	35.75	52.38	35.25	0.62	0.43	0.42	0.47
GPT-4o	64.6	39.60	67.00	51.8	55	43.20	62.22	54.4	0.65	0.47		0.49
Qwen-VL-chat	14.04	4.51	9.375	12	9.58	7.14	35.23	8.4	0.49	0.33	0.31	0.36
CogAgent-VQA	14.89	5.95	11.28	9.4	13.07	11.52	19.36	8.8	0.45	0.29	0.15	0.11
Intern-VLM-4khd	26.67	13.87	22.22	17.2	28.71	18	29.84	9.6	0.52	0.36	0.32	0.34
Interleaved Image-text Baseline												
Gemini-1.5 Flash	60.42	33.33	50.44	50.39	53.22	40.17	62.90	48.02	0.52	0.42	0.42	0.51
GPT-4o	72.47	49.26	69.6	47.6	66.5	48.93	57.77	54	0.56	0.51	0.46	0.49
Qwen-VL-chat	12.86	6.64	11.61	10.29	9.59	5.38	12.88	7.09	0.16	0.17	0.05	0.09
Idefics-Mantis	10.46	2.62	10.39	8.49	2.8	5.69	9.09	3.61	0.34	0.22	0.30	0.3

Table 3: Results on sampled subset of MMTabQA. Substring match is reported for Wiki-realted data sources and ROUGE-L is reported for FetaQA data source. EQ - Explicit Questions, AQ - Answer-Mention Questions, IQ - Implicit Questions, VQ - Visual Questions. Best performing models are highlighted in red.

6.1 Performance across Strategies

As described in Section 4, the Partial Input Baseline forms the lower bound for our experiments (Table 3). Moreover, the Oracle-Entity Replaced Baseline establishes the experimental upper bound, showcasing superior performance compared to all other baselines. This baseline reflects the model's performance under ideal conditions where entity disambiguation is executed with 100% accuracy.

The Table-as-image baseline integrates missing image information beyond what the Partial Input baseline provides, resulting in an expected improvement in performance. However, the challenge of interpreting table structure directly from the image remains evident for LLMs, thus leading to consistent performance limitations.

We extend upon this with the interleaved Image-text baseline, employing separate encoding for text and images, which yields enhanced representations and improved performance relative to the Table-as-image baseline. However, it is noted that this performance does not achieve parity with our defined Upper Bound, highlighting avenues for further enhancement. We additionally explore the image-as-caption approach, aiming to encode multimodal information into textual form for QA. Our findings indicate that while this method is less effective than the interleaved Image-text baseline, it outperforms the table-as-image baseline. This underscores the persistent challenge of accurately interpreting table structure & text from a singular image.

6.2 Performance across Models

Closed-source models like GPT-4o and Gemini-1.5 Flash outperform open-source models in multimodal tasks due to advanced training techniques and better integration of visual and textual data. In text-only tasks, the performance gap between open-source and closed-source models narrows significantly, with open-source models like Llama-3 providing competitive results.

Overall, we see that closed-source models generally outperform open-source models. GPT-4o demonstrates the best performance, achieving a substring match as high as 60.6% and 42.6% for WikiTableQuestion Explicit and Answer-mention Questions in the interleaved images approach. This is closely followed by Gemini-1.5 Flash, which achieves a substring match of up to 61% and 33.33% for the same dataset subsection. We also note that the true reasoning capabilities of GPT-4o might be more advanced when provided with Interleaved input, as it refuses to answer some questions due to policy violations.

Notably, open-source models provide competitive results in text-only baselines. Here, the performance gap between is around 10% to 20%, which indicates that open-source textual LLMs with a large number of parameters are competitive with state-of-the-art closed-source textual LLMs. While the performance of Llama-3 is on-par, the performance of Mixtral 8x7B lags behind. This is because with 9X more parameters, Llama-3 is capable of much more complex reasoning and parametric knowledge than Mixtral8x7B. Furthermore, the Partial Input baseline demonstrates that Open Source models leverage real-world knowledge to infer missing entities.

Their performance notably declines in multimodal baselines, particularly in approaches like Table as an Image and Interleaved Text-Image. In Vision-Language models, the disparity between Open-Source and Closed-Source models becomes more pronounced. Table-as-image models encounter challenges such as entity disambiguation within tables, highlighting deficiencies in parametric multimodal knowledge and table structure parsing, which reflects their relatively weaker Vision Encoders. Similarly, in interleaved models, Open-Source counterparts struggle to contextualize multiple images, often resulting in nonsensical answers influenced predominantly by one image rather than considering all provided images.

6.3 Performance across Data Sources

We examine Table 3 to gain deeper insights into our proposed tasks and model performance. We observe that the performance of models is similar on the WikiTableQuestions dataset and the WikiSQL datasets, since both are short-form question datasets and require a similar kind of entity disambiguation as a challenge for the question-answering. On FeTaQA, we notice that the ROUGE-L scores themselves don't vary much between Upper Bound and Lower Bound. This is because majority of the N-grams used for computing the metric wouldn't involve the image-replaced entity. We observe a significant decrease in scores of Image-Captioning baseline on FeTaQA dataset. This decline is likely due to the inclusion of text from provided captions and visual descriptions, which adversely affected recall on the gold summaries. In addition, we observe slight variations in ROUGE scores, which offer only a rough indication of VLM performance.

6.4 Performance across Question types

Models typically exhibit superior performance in scenarios involving simple reasoning. Conversely, tasks requiring complex reasoning or multi-step inference frequently lead to model failures. This observed trend underscores the challenges faced by current models in handling intricate reasoning processes. Explicit questions, as defined in the preceding discussion, contain clear and specific entity mentions within the query, facilitating their resolution by computational models. This assertion is substantiated by the data presented in Table 3, where explicit questions achieve the highest parameter scores in comparison to other question types.

Implicit questions, in contrast, necessitate additional reasoning to infer the answers, resulting in lower model performance relative to explicit questions. When analyzing the WikiSQL dataset, implicit questions achieve a higher performance metric (52.3%) compared to explicit questions (47.08%). This discrepancy can be attributed to the nature of reasoning required by each question type. For tasks requiring very complex reasoning, as seen in WikiTableQuestions (WTQ), explicit questions tend to perform better. However, in the context of WikiSQL, where reasoning is primarily simple and SQL-based, implicit questions exhibit superior performance. This trend is corroborated by the oracle-baseline performance observed on both WikiSQL and WTQ datasets.

Explicit Answer Mention Questions pose a significant challenge as they require the generation of answers that mention entities replaced by images. This task demands precise entity disambiguation, leading to lower performance metrics across all data sources. Specifically, analyzing Gemini's performance in the "Table as an Image" approach, there is a marked decrease in performance—over

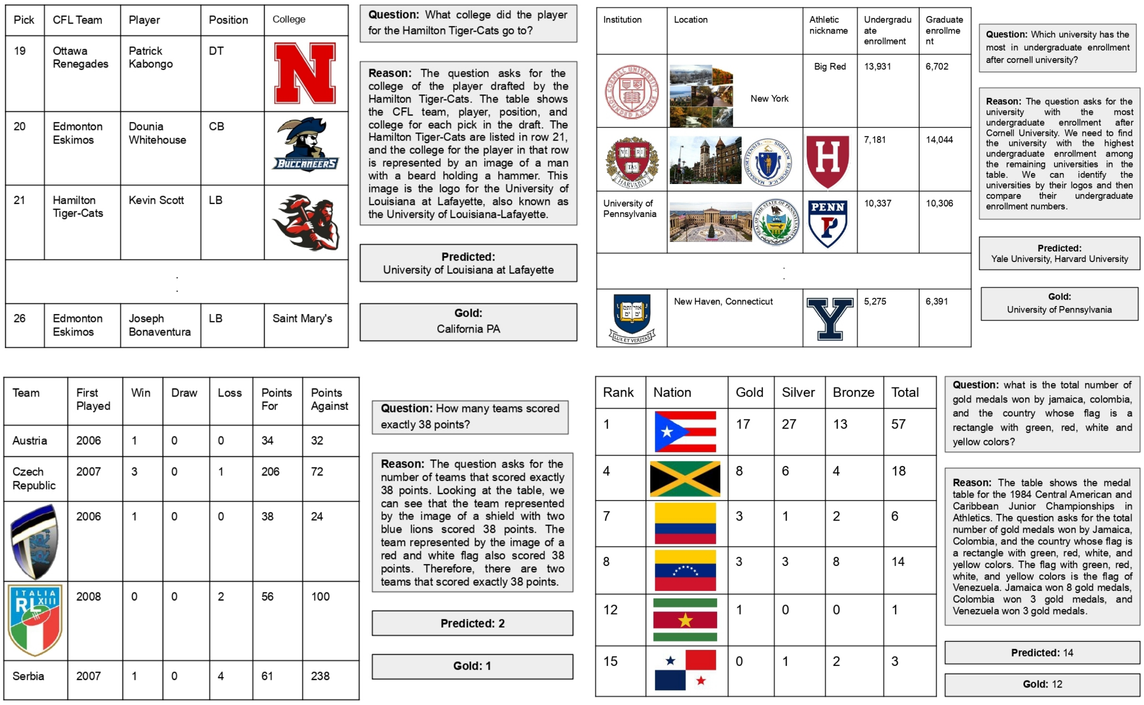


Figure 3: Clockwise from left - (a): College Football Table illustrating Entity Disambiguation Problem, (b): College Enrollment Table illustrating Complex Reasoning Problem, (c): 1984 Central American Games Table illustrating Visual Attribute Problem, (d): International Football Table illustrating Excessive Content Handling Problem.

10%—for answer mention questions compared to other question types.

Visual questions perform better than answer-mention questions but underscore a significant limitation of current language models. These questions necessitate a visual understanding of entities depicted in images, a task that is considerably more challenging than leveraging images for explicit or implicit reasoning. This challenge is reflected in the performance metrics: 37.8% on WikiSQL and 35.25% on WikiTableQuestions using the "Table as an Image" approach. While the FeTaQA dataset presents valuable information, its current metrics limit the extent of inferences we can draw.

7 What did we learn?

A significant observation from our experiments is the varied performance of models across different baseline settings and the reasons underlying these disparities. GPT-4o demonstrates competitive performance across all data sources and methodologies and those results are used for our analysis. We list some primary issues below:

Identification of Visual Attributes : This presents challenges for multimodal models, particularly in recognizing crucial visual elements within an image required to answer associated questions. For example, in Figure 3(c), the model fails to correctly identify a flag based on its colors.

Entity Disambiguation : The model inaccurately identifies entities from images, leading to errors. For instance, in Figure 3(a), the model misidentifies the logo of California PA as that of the University of Lafayette.

Handling Excessive Content : Challenges arise in handling excessive content, leading to instances of incomplete or incorrect retrieval by the model. Figure 3(d) illustrates such a scenario where the model's incomplete comprehension of the table leads to erroneous conclusions.

Incorrect Conclusions : Despite correctly identifying entities at times, the model occasionally reaches incorrect conclusions, possibly due to incorrect reasoning, resulting in erroneous answers, as depicted in Figure 3(b). These findings highlight the weaknesses of VLMs in image handling, particularly concerning the capabilities of their vision encoders. Notably, open-source models underperform compared to closed-source models as image complexity increases, both in intricacies (e.g., the table-as-an-image approach) and in quantity (e.g., the interleaved-image approach).

Open-source VLMs often lack vision encoders capable of handling intricate or multiple images, resulting in inaccurate interpretations. Moreover, even when some level of image interpretation is achieved, the limited reasoning abilities of these models render them highly ineffective for comprehensive analysis. Additionally, we also perform a quantitative analysis of these errors on 720 randomly sampled incorrect responses of GPT-4o based on the Table-as-image and Interleaved input approaches on broader error categories. The quantitative insights are given below:

Type	WikiSQL	WikiTable	FetaQA
1	32.17%	34.48%	22.08%
2	3.04%	2.58%	0.83%
3	54.34%	44.82%	63.75%
4	10.43%	18.10%	13.33%

Table 4: Error Analysis - A dataset perspective. Type 1: Entity Disambiguation Issues, Type 2: Context Length Related Issues, Type 3: Reasoning and Text Input Errors, Type 4: Identification of Visual Attributes. Further description present in Appendix F

Type	EQ	AQ	IQ	VQ
1	28.65%	43.57%	30.81%	15.00%
2	1.75%	0.55%	2.32%	3.88%
3	64.91%	48.04%	61.05%	44.44%
4	4.67%	7.82%	5.81%	36.67%

Table 5: Error Analysis - A question type perspective. Type 1: Entity Disambiguation Issues, Type 2: Context Length Related Issues, Type 3: Reasoning and Text Input Errors, Type 4: Identification of Visual Attributes. Further description present in Appendix F

Our analysis reveals that the majority of errors can be attributed to reasoning errors, entity disambiguation issues, and difficulties with visual aspect identification, in that order. These findings highlight the current models' inability to effectively process multimodal table data for QA purposes, thereby reinforcing the necessity of our dataset.

8 Comparison with Related Work

Recent advancements in natural language processing (NLP) have expanded beyond traditional homogeneous tables to incorporate additional modalities. Works such as (Chen et al., 2020; Zhu et al., 2021; Chen et al., 2021; Zhao et al., 2022) integrate paragraph context alongside tables for enhanced tabular question answering. Meanwhile, efforts like (Talmor et al., 2021; Li et al., 2022) introduce images alongside text and tables, yet they do not address the non-homogeneous modalities found in MMTabQA tables, containing text-only tables.

Visual Table understanding has also gained attention, with approaches such as (Zheng et al., 2024; Kim et al., 2024) converting textual tables into visual formats for multimodal reasoning. However, these efforts do not capture the multimodal complexities inherent in MMTabQA. Datasets like (Marino et al., 2019; Schwenk et al., 2022) focus on general knowledge for Visual Question Answering, while others (Lerner et al., 2022; Mensink et al., 2023; Chen et al., 2023) emphasize fine-grained entity knowledge within images. Multimodal Entity Linking and Disambiguation across different modalities are explored in (Moon et al., 2018; Wang et al., 2022), echoing the entity linking challenges posed by MMTabQA. In Visual Question Answering, advancements have been made in handling multiple images (Penamakuri et al., 2023; Jiang et al., 2024b), though it remains less explored compared to single-image tasks.

Unlike prior work, which typically focuses on either textual or visual elements separately, our task confronts the novel challenge of multimodal tables that integrate multiple texts and multiple images within table cells. This involves addressing explicit, implicit, and answer-mention questions, while also advancing visual understanding within the framework of table-based reasoning.

9 Conclusion

This research explores whether NLP models can effectively reason with knowledge on multimodal structured data. We investigate their ability to process tables that combine images and text, introducing MMTabQA, a new dataset for this task. Our experiments reveal substantial challenges for AI models in integrating and interpreting multiple text and image inputs, understanding visual context, and comparing visual content across images. Our findings position MMTabQA as a crucial benchmark for advancing AI's capabilities in analyzing multimodal structured data.

Future Directions. Our research presents opportunities for expansion by enhancing existing Wikipedia-derived datasets through augmentation and proposing a human-annotated dataset from real-world multimodal tables beyond Wikipedia. Diversifying with additional datasets will enrich our dataset's diversity and scope. Addressing model errors during retrieval is a significant challenge, tackled through Retrieval-Augmented Generation (RAG). Moreover, optimizing open-source models tailored to our task is crucial, focusing on efficient models capable of achieving results comparable to computationally intensive counterparts. These efforts aim to advance multimodal table reasoning.

Limitations

Our work has several notable limitations. Chiefly, financial and computational resource constraints prevented us from fine-tuning all the models considered, potentially underrepresenting their capabilities beyond our primary focus. Additionally, the language limitations in this research, particularly the emphasis on English for creating Multimodal Reasoning datasets and methodologies, highlight the necessity of linguistic diversity in NLP applications to ensure broader applicability and inclusivity. Considering the novelty of the task, it is also important to recognize that our insights may not be exhaustive, pointing to the potential for future research.

Ethical Statement

As the work's authors, we certify that our investigation and publication adhere to the strictest ethical guidelines. For the purpose of making our results more reproducible, we include comprehensive information which involves disclosing code, datasets (we work with publicly accessible datasets and adhere to the ethical guidelines established by the datasets' creators), and other pertinent materials. The dataset in this study is designed for research on multimodal table question answering. It should be strictly used for research purposes, not for other applications As a result, the scientific community can verify and build upon our findings. The assertions made in this paper align with the outcomes of our experiments. But because black-box big language models are inherently stochastic, we have reduced variability by keeping the temperature constant. In order to ensure the reproducibility of our work, we provide comprehensive details about the prompting techniques utilized, models used, dataset splits, and annotations made.

Acknowledgement

Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-20-1-0080. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This work was partially funded by ONR Contract N00014-19-1-2620. Lastly, we extend our appreciation to the reviewing team for their insightful comments.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

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.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

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.

Ana Cláudia Akemi Matsuki de Faria, Felype de Castro Bastos, José Victor Nogueira Alves da Silva, Vitor Lopes Fabris, Valeska de Sousa Uchoa, Décio Gonçalves de Aguiar Neto, and Claudio Filipi Goncalves dos Santos. 2023. Visual question answering: A survey on techniques and common trends in recent literature. *arXiv preprint arXiv:2305.11033*.

Xi Fang, Weijie Xu, Fiona Anting Tan, Jiani Zhang, Ziqing Hu, Yanjun Qi, Scott Nickleach, Diego Socolinsky, Srinivasan Sengamedu, and Christos Faloutsos. 2024. Large language models(llms) on tabular data: Prediction, generation, and understanding – a survey. *Preprint*, arXiv:2402.17944.

Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, et al. 2023. Cogagent: A visual language model for gui agents. *arXiv preprint arXiv:2312.08914*.

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, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024a. Mixtral of experts. *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. *Preprint*, arXiv:2405.01483.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: Recent advances. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, pages 174–186, Singapore. Springer Nature Singapore.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3108–3120.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2023. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *International Conference on Learning Representations (ICLR)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. *Preprint*, arXiv:1906.00067.

T. Mensink, J. Uijlings, L. Castrejon, A. Goel, F. Cadar, H. Zhou, F. Sha, A. Araujo, and V. Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3090–3101, Los Alamitos, CA, USA. IEEE Computer Society.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Thomas Müller, Julian Eisenschlos, and Syrine Krichene. 2021. TAPAS at SemEval-2021 task 9: Reasoning over tables with intermediate pre-training. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 423–430, Online. 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.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Abhirama Subramanyam Penamakuri, Manish Gupta, Mithun Das Gupta, and Anand Mishra. 2023. Answer mining from a pool of images: towards retrieval-based visual question answering. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI '23.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*.

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*, abs/2104.06039.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Changyao Tian, Xizhou Zhu, Yuwen Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Yuntao Chen, Lewei Lu, Tong Lu, Jie Zhou, Hongsheng Li, Yu Qiao, and Jifeng Dai. 2024. Mm-interleaved: Interleaved image-text generative modeling via multi-modal feature synchronizer. *Preprint*, arXiv:2401.10208.

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

Nancy X. R. Wang, Diwakar Mahajan, Marina Danilevsky, and Sara Rosenthal. 2021. SemEval-2021 task 9: Fact verification and evidence finding for tabular data in scientific documents (SEM-TAB-FACTS). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 317–326, Online. Association for Computational Linguistics.

Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. *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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

A Dataset Statistics

Table 6 highlights the complexity of reasoning involved in different questions based upon the number of columns involved. We can see FeTaQA involving higher multi-column reasoning questions.

Data Source	Single col. reasoning	Multi col. reasoning
WikiSQL	17,558	3,914
WikiTable-Questions	8,952	1,100
FeTaQA	4,620	2,856
HybridQA	26,358	4,112

Table 6: MMTabQA Reasoning Complexity

Table 7 presents the different types of questions in MMTabQA, each presenting a different type of challenge. While explicit questions require disambiguating the entity mentioned in the question, answer-mention questions are more complex because they need to generate an image-replaced entity in the answer. Implicit questions on the other hand involve more logical reasoning, while visual questions require the model to understand the visual aspects of images in the table specifically.

Table 8 highlights the top 10 domains of the tables in the different datasets, indicating the topics on which the tables and the corresponding questions are based on. Table 9 provides an overview of image category distributions across four prominent

Data Source	Explicit Ques	Implicit Ques	Visual Ques	Answer-Mention Ques
FeTaQA	2,499	612	1,185	3,180
WikiTable-Questions	3,523	2,879	877	2,773
WikiSQL	12,956	315	1,827	6,374
HybridQA	5,819	17,647	1,874	5,130

Table 7: Question Type Statistics for MMTabQA

Domain (%)	WTQ	FetaQA	WikiSQL	HybQA
STEM	30.18	29.00	28.87	29.28
Media	16.04	17.73	16.49	17.75
Biography	14.05	15.20	14.23	13.85
None	11.83	11.40	11.38	10.92
Europe	11.43	12.40	12.23	12.50
NA	8.26	10.24	9.20	9.34
P&G	6.98	5.91	5.96	5.36
Technology	6.75	6.52	6.69	6.90
Asia	6.51	7.20	6.91	6.86
P&R	5.24	-	-	-
Literature	-	4.83	4.63	4.81

Table 8: Top 10 Domains of the tables categorized based on the topic of the Wikipedia page. P&G: Politics and Government, P&R: Philosophy and Religion, WTQ: WikiTableQuestions, HybQA: HybridQA, NA: North America.

datasets. The analysis reveals a consistent emphasis on human and logo categories across all datasets, indicating these entities are central to the types of questions posed. Beyond humans and logos, there exists notable variability in other categories such as location/landscapes, seals, coat of arms, flags, and posters across the datasets.

Table 10 shows the distribution of categories of the entities in the answer, highlighting the specific entities based upon which the questions are posed in the dataset. Some additional statistics on the dataset are presented in table 11.

B Dataset Examples

To demonstrate the quality and features of our created MMTabQA we provide table examples along with question - answer pair from all the four datasets. Fig. 4 shows the examples from WikiTableQuestions dataset, Fig. 5 shows the examples from WikiSQL dataset, Fig. 6 shows the examples from FeTaQA dataset, Fig. 7 shows the examples from HybridQA dataset. Each example depicts the Multi-modal table along with a Question - Answer sample.

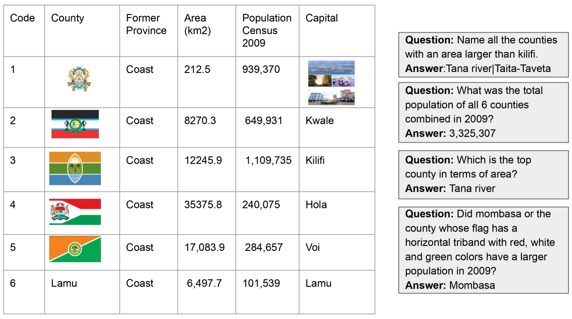


Figure 4: WikiTableQuestions Dataset Example

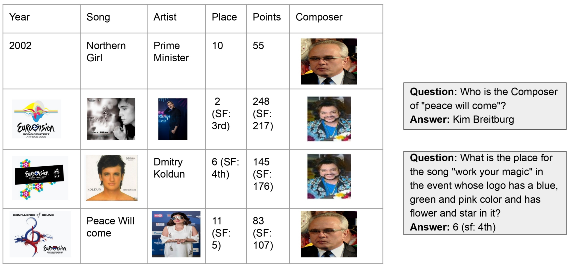


Figure 5: WikiSQL Dataset Example

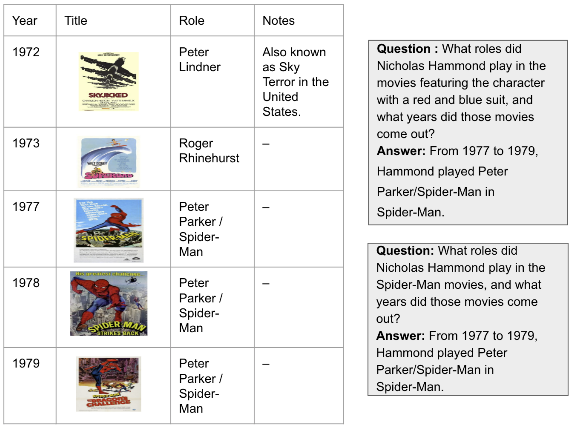


Figure 6: FeTaQA Dataset Example

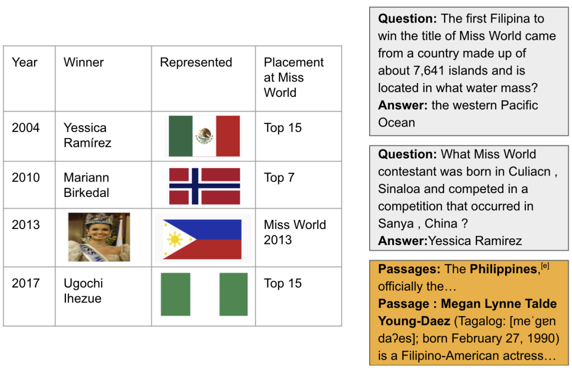


Figure 7: HybridQA Dataset Example

Dataset	Human	Location/ landscapes	Seals	Coat of Arms	Flags	Poster	Logo	Miscellaneous
WikiTable-Questions	6,305	3,082	356	460	831	455	2,380	1,518
FetaQA	10,043	3,779	478	779	1,158	5,446	5,628	8,372
WikiSQL	16,915	4,518	738	703	1,149	751	4,572	5,856
HybridQA	31,816	4,219	868	2,193	2,053	2,313	7,794	11,090

Table 9: Image Category Distribution in MMTabQA

Dataset	Human	Location	Product	Time	Money	Event	Number	Org.	Boolean	Other
WikiTable-Questions	1,639	1,598	628	430	43	117	3,947	901	210	539
FetaQA	957	493	1,410	1,320	33	173	1,792	428	99	771
WikiSQL	3,491	3,721	649	2,026	1,866	112	4,254	2,623	24	4,386
HybridQA	4,791	5,276	2,093	4,838	296	299	5,079	3,382	3	4,326

Table 10: Distribution of answer entity categories

Dataset	Avg. No. of Rows	Questions per Table	Avg. No. of Cols	Total Img Occur.	Total Unique Images	Avg. Unique Images	Avg. Images per Table
WikiTable-Questions	18.23	7.98	6.27	32,304	16,338	17.67	25.66
FetaQA	14.44	1.27	6.12	102,785	37,238	10.43	17.43
WikiSQL	13.95	2.19	6.25	207,343	37,354	13.68	21.19
HybridQA	15.87	3.77	4.50	153,246	62,346	14.64	18.95

Table 11: Additional MMTabQA Statistics. Org. stand for Organization.

C HybridQA Experiments

In addition to the data sources analyzed previously, we have incorporated the HybridQA dataset to enhance our proposed task with a question-answering component that requires reasoning over heterogeneous information. HybridQA aligns each question with a Wikipedia table and multiple free-form text corpora linked to the entities within the table. The design of the questions necessitates the aggregation of both tabular and textual information, rendering them unanswerable if either form is lacking. Oracle retrieval is employed to obtain the relevant passages for question-answering tasks.

We benchmark our augmented dataset utilizing the four approaches outlined earlier: text-only baselines (Partial Output Baseline and Oracle Entity Baseline), the Table-as-an-Image approach, and the Interleaved Text-Image approach, as presented in Table 12. Exact Match, Substring Match, and F-1 Score are the metrics employed to evaluate the model's results. For the purpose of analysis, we will primarily focus on Substring Match.

Examining different approaches, Llama 3-70B and Gemini-1.5 Flash demonstrate comparable performance on text-only baseline models, indicating that the open-source model is equally capable as the closed-source model. Mistral 8x7B, however, underperforms, which can be attributed to the fewer parameters it contains.

In multimodal baselines, GPT-4o exhibits the best performance, with Gemini-1.5 Flash being a close second for both Table as Image and Interleaved Text-Image approaches. Open-source models display an interesting trend for these tasks. In the Table as Image approach, CogAgent-VQA and Intern-VLM-xcomposer-4khd provide decent performance, comparable to closed-source models, whereas Qwen-VL seems to underperform, likely due to the same parameter-related issues faced by Mistral in the text-only baseline.

A clear distinction emerges for the Interleaved Text-Image approach: closed-source models outperform open-source models, with GPT-4o being the best. Open-source models struggle to handle and infer from multiple images, and their smaller size further limits their performance.

A major observation is that for both text-only approaches, which represent the boundary values for the task, the performance metrics are quite close. Overall, the models demonstrate decent per-

Models	EQ			AQ			IQ			VQ		
Metrics	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1
Oracle-Entity Replaced Baseline												
Gemini 1.5 Flash	63.71	77.14	0.77	46.49	53.65	0.54	60.83	71.59	0.73	-	-	-
GPT-4o	31.00	84	0.47	28.85	73.71	0.41	26.28	81.28	0.41	-	-	-
LLAMA 3 70B	61.76	75.47	0.74	55.25	61.58	0.62	60.86	73.09	0.72	-	-	-
Mixtral 8x7B	46.00	63.90	0.63	34.97	53.90	0.45	33.67	67.63	0.50	-	-	-
Partial Input Baseline												
Gemini 1.5 Flash	59.71	71.57	0.71	28.71	32.43	0.34	59.14	69.71	0.70	43.29	53.46	0.54
GPT-4o	39.57	79.42	0.55	28.42	49.42	0.37	44.14	78.85	0.57	28.25	65.65	0.41
LLAMA 3 70B	55.56	70.13	0.68	34.01	39.91	0.41	56.83	68.63	0.69	45.04	59.92	0.57
Mixtral 8x7B	40.23	58.16	0.56	26.01	30.92	0.33	45.44	59.04	0.58	28.90	41.37	0.42
Image-captioning Baseline												
Gemini-1.5 Flash	59.57	71.85	0.71	39.57	43.00	0.46	59.14	67.28	0.69	44.91	52.64	0.55
Table-as-an-Image												
Gemini 1.5 flash	48.50	67.67	0.64	26.14	33.28	0.34	47.78	68.10	0.63	42.07	58.54	0.56
GPT-4o	62.05	76.31	0.75	48.00	53.00	0.56	64.46	75.50	0.76	50.81	61.18	0.63
Qwen-VL-chat	12.81	16.08	0.17	7.31	10.32	0.13	9.87	13.16	0.14	6.72	11.00	0.11
CogAgent-VQA	15.74	19.64	0.211	9.58	11.79	0.15	15.29	20.00	0.21	11.39	14.55	0.17
Intern-VLM-4khd	43.60	59.74	0.57	22.25	26.59	0.29	43.95	57.10	0.56	35.48	44.81	0.46
Interleaved Text-Image Baseline												
Gemini 1.5 Flash	45.94	71.21	0.63	25.77	36.79	0.35	49.13	74.05	0.66	39.76	60.00	0.57
GPT4o	59.80	80.20	0.76	43.40	49.80	0.52	51.80	69.20	0.68	45.53	60.77	0.61
Qwen-VL	1.66	5.80	0.03	2.12	7.43	0.05	1.64	10.27	0.03	1.03	4.52	0.02
Idefics-Mantis	0.58	2.05	0.02	0.00	0.92	0.02	1.11	3.78	0.02	0.54	1.34	0.01

Table 12: Detailed results on sampled subset of HybridQA. EM - Exact Match, SSM - Substring Match, F1 - F1 score. EQ - Explicit Questions, AQ - Answer-Mention Questions, IQ - Implicit Questions, VQ - Visual Questions. Best performing models are highlighted in red.

formance across all tasks and baselines. This can be attributed to the use of passages as additional context, which facilitates entity disambiguation for the models. However, this approach undermines the primary objective of our task, which is to challenge the models' ability to reason over heterogeneous information without relying heavily on supplementary textual context.

D Additional Metrics

We present a detailed benchmark report for our dataset across various data sources. For WikiSQL (Fig. 14) and WikiTableQuestions (Fig. 13), we report Exact Match (EM), Substring Match (SSM), and F1 Score. For FeTaQA (Fig. 15), we include BLEU Score, ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L). These detailed evaluations provide a thorough understanding of the models' capabilities and performance variations across different datasets, highlighting their strengths & weakness.

E Visual Question Error Analysis

Upon performing a further fine grained analysis of the incorrect marked samples we identify 5 types of errors in the visual questions (breakdown of the 15% questions) and suggest some ways of rectifying them in future:

Non-repairable (0.6%) – These are visual questions based on portraits/paintings where no unique visual attributes are present in the table. We can identify them by prompting a VLM to check for visual attributes present in multiple images and discard them if found.

Hallucinated attributes (6.6%) – These questions have partially or completely incorrect visual attributes for the entity. We can use a VLM to check if the attributes are present in the entity's image and discard them if they aren't.

Question-Type	EQ			IQ			AQ			VQ		
Metrics	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1
Partial Input Baseline												
Gemini 1.5 Flash	38.16	40.99	0.42	25.05	27.39	0.35	46.70	48.96	0.48	28.40	31.40	0.31
GPT-4o	52.19	57.44	0.55	33.87	38.01	0.43	70.83	73.43.54	0.72	39.00	42.4	0.42
LLAMA 3 70B	39.43	41.13	0.43	24.50	26.49	0.32	41.49	43.75	0.43	29.40	31.80	0.34
Mixtral 8x7B	23.58	26.56	0.26	8.83	9.91	0.13	28.17	30.26	0.30	17.60	20.20	0.20
Oracle-Entity Replaced Baseline												
Gemini 1.5 Flash	71.21	74.89	0.74	75.68	78.20	0.83	53.82	54.86	0.55	-	-	-
GPT-4o	52.62	87.80	0.59	55.85	84.86	0.66	50.69	84.54	0.55	-	-	-
LLAMA 3.00 70B	53.62	75.74	0.61	53.69	75.32	0.67	45.49	58.85	0.49	-	-	-
Mixtral 8x7B	48.79	54.89	0.54	48.47	53.87	0.57	37.57	40.70	0.40	-	-	-
Image Captioning Baseline												
Gemini 1.5 Flash	48.65	52.34	0.52	34.41	42.16	0.48	48.96	51.39	0.50	38.20	42.20	0.43
Table-as-an-image baseline												
Gemini 1.5 Flash	40.80	44.22	0.44	22.30	25.65	0.34	37.35	41.01	0.39	33.20	37.80	0.38
GPT-4o	60.80	64.60	0.66	36.20	39.60	0.49	65.40	67.00	0.67	50.20	51.80	0.50
Qwen-VL-chat	12.20	14.04	0.14	3.60	4.50	0.07	7.99	9.38	0.09	10.40	12.00	0.13
CogAgent-VQA	12.62	14.89	0.15	5.59	5.95	0.10	8.68	11.28	0.10	8.40	9.40	0.09
Intern-VLM-4khd	22.55	26.67	0.27	11.35	13.87	0.20	18.58	22.22	0.22	15.40	17.20	0.18
Interleaved Image-text Baseline												
Gemini 1.5 Flash	47.98	60.31	0.53	20.33	33.33	0.36	44.21	50.45	0.47	38.58	50.39	0.45
GPT4o	69.65	72.47	0.72	44.44	49.27	0.57	68	69.6	0.69	46.40	47.60	0.49
Qwen-VL	1.04	12.87	0.02	0.00	6.65	0.02	0.19	11.62	0.01	0.42	10.29	0.02
Idefics-Mantis	5.44	10.46	0.07	2.33	2.62	0.04	3.46	10.39	0.06	3.77	8.49	0.05

Table 13: Detailed results on sampled subset of WikiTableQuestions. EM - Exact Match, SSM - Substring Match, F1 - F1 score. EQ - Explicit Questions, AQ - Answer-Mention Questions, IQ - Implicit Questions, VQ - Visual Questions. Best performing models are highlighted in red.

Hard to identify attributes (1.8%) – These questions rely on visual attributes that are hard to spot in the table image and require a zoomed-in, high-resolution view. While technically correct, they aren't relevant for table question answering. A VLM can help identify and prune these by checking if the attributes are easily noticeable in the key entity's image.

Non-unique attributes (6.4%) – These questions involve non-unique visual attributes for the entity, but unlike non-repairable questions, a unique set of attributes is possible. We can identify them by using a VLM to check if the attribute appears in multiple images and discard those that do.

No visual attribute (0.6%) – These questions refer directly to the entity name, as in a logo or poster (e.g., "Star Wars" on a Star Wars poster). They can be filtered by checking if the question's tokens completely overlap with the entity name.

F Error Analysis Labels

Our label classification is as follows:

1. Entity Disambiguation Issues: Instances where the model fails to accurately identify the entity mentioned in the question, leading to incorrect interpretations (Fig 3a).
2. Context Length-Related Issues: Cases where the model struggles to comprehend prompts due to lengthy context or multiple images, resulting in incorrect or no output.
3. Reasoning and Text Input Errors: Situations where the model's final output is incorrect due to faulty table interpretation, erroneous information extraction (Fig 3c), model hallucination, or incorrect reasoning (Fig 3b).
4. Visual Attribute Identification Errors: Instances where the model incorrectly identifies visual aspects of an image, leading to erroneous answers (Fig 3d).

Question-Type	EQ			IQ			AQ			VQ		
Metrics	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1	EM	SSM	F1
Partial Input Baseline												
Gemini 1.5 Flash	36.43	39.14	0.37	26.71	28.71	0.38	57.78	62.22	0.21	24.60	28.00	0.30
GPT-4o	49.42	52.57	0.48	39.14	43.85	0.56	66.98	72.38	0.67	33.6	39.0	0.39
LLAMA 3 70B	38.25	41.12	0.38	27.75	30.76	0.41	56.83	61.27	0.57	27.80	30.60	0.33
Mixtral 8x7B	20.86	23.43	0.22	13.86	17.71	0.24	24.13	28.89	0.25	15.00	19.20	0.21
Oracle-Entity Replaced Baseline												
Gemini 1.5 Flash	79.00	82.29	0.73	80.57	81.86	0.90	73.02	77.46	0.72	-	-	-
GPT-4o	81.85	85.57	0.75	81.57	82.71	0.91	73.01	79.04	0.72	-	-	-
LLAMA 3.00 70B	74.29	78.29	0.70	77.00	78.57	0.88	62.86	68.25	0.63	-	-	-
Mixtral 8x7B	54.71	59.29	0.55	60.71	65.29	0.75	27.94	33.97	0.28	-	-	-
Image Captioning Baseline												
Gemini 1.5 Flash	45.43	50.43	0.45	33.71	40.86	0.53	62.54	67.30	0.65	40.80	46.60	0.45
Table-as-an-image baseline												
Gemini 1.5 flash	43.63	47.08	0.43	28.27	35.75	0.47	46.03	52.38	0.49	32.51	35.25	0.35
GPT-4o	51.60	55.00	0.48	39.00	43.20	0.59	58.73	62.22	0.60	47.80	54.40	0.52
Qwen-VL-chat	6.29	9.59	0.10	6.20	7.14	0.12	17.14	35.24	0.19	4.60	8.40	0.08
CogAgent-VQA	10.07	13.08	0.14	10.27	11.53	0.16	6.98	19.37	0.07	5.80	8.80	0.09
Intern-VLM-4khd	24.00	28.71	0.26	15.14	18.00	0.27	23.49	29.84	0.26	5.80	9.60	0.11
Interleaved Image-text Baseline												
Gemini 1.5 - Flash	56.95	53.22	0.54	32.59	40.18	0.49	53.23	62.90	0.54	43.61	48.02	0.49
GPT4o	63.00	66.50	0.62	39.96	48.93	0.61	55.24	57.78	0.56	48.60	54.00	0.49
Qwen-VL	2.63	9.60	0.05	1.82	5.38	0.04	5.08	12.88	0.06	1.14	7.09	0.03
Idefics-Mantis	1.33	2.88	0.03	4.85	5.70	0.08	0.00	9.09	0.04	1.08	3.61	0.03

Table 14: Detailed results on sampled subset of WikiSQL. EM - Exact Match, SSM - Substring Match, F1 - F1 score. EQ - Explicit Questions, AQ - Answer-Mention Questions, IQ - Implicit Questions, VQ - Visual Questions. Best performing models are highlighted in red.

G Prompt Samples

We provide a detailed sample for each baseline strategy to illustrate our approach: Partial Input (Fig. 8), Oracle-Based Entity (Fig. 9), Image-Caption Fig.(10), Table-Image (Fig. 11), and Interleaved Text-Image (Fig. 12). These examples demonstrate the varying degrees of information and context provided to the models, highlighting the differences in their ability to process and respond to diverse types of input. Through these samples, we aim to showcase the challenges and nuances involved in each strategy, offering insights into the models' performance.

Models	EQ				AQ				IQ				VQ			
Metrics	BLEU	R-1	R-2	R-L	BLEU	R-1	R-2	R-L	BLEU	R-1	R-2	R-L	BLEU	R-1	R-2	R-L
Oracle-Entity Replaced Baseline																
Gemini 1.5 Flash	29.93	0.62	0.42	0.53	23.29	0.62	0.39	0.50	16.37	0.50	0.28	0.41	-	-	-	-
GPT4o	24.67	0.64	0.43	0.53	18.88	0.60	0.37	0.48	15.78	0.52	0.31	0.43	-	-	-	-
LLAMA 3 70B	29.13	0.66	0.45	0.54	19.23	0.58	0.36	0.46	15.10	0.50	0.28	0.41	-	-	-	-
Mixtral 8x7B	7.97	0.49	0.31	0.42	11.16	0.51	0.31	0.41	7.54	0.41	0.22	0.33	-	-	-	-
Partial Input Baseline																
Gemini 1.5 Flash	16.24	0.52	0.33	0.44	20.17	0.54	0.31	0.44	20.39	0.53	0.31	0.44	22.89	0.55	0.34	0.47
GPT4o	28.21	0.61	0.40	0.50	20.51	0.56	0.34	0.46	19.61	0.51	0.29	0.42	16.69	0.52	0.32	0.44
LLAMA 3 70B	31.49	0.63	0.42	0.53	21.88	0.56	0.33	0.46	19.69	0.54	0.31	0.45	24.42	0.56	0.35	0.48
Mixtral 8x7B	18.39	0.54	0.34	0.44	12.97	0.48	0.26	0.39	11.73	0.47	0.26	0.38	14.11	0.48	0.27	0.39
Image-Captioning Baseline																
Gemini 1.5 flash	4.31	0.57	0.37	0.48	7.43	0.56	0.34	0.46	8.19	0.50	0.28	0.42	7.52	0.51	0.31	0.43
Table-as-an-Image																
Gemini 1.5 flash	29.93	0.62	0.42	0.53	18.73	0.54	0.32	0.44	7.09	0.51	0.30	0.43	13.46	0.56	0.34	0.47
GPT-4o	29.13	0.66	0.45	0.54	19.74	0.58	0.35	0.47	19.08	0.55	0.33	0.45	21.70	0.59	0.37	0.49
Qwen-VL-chat	7.97	0.49	0.31	0.42	5.21	0.40	0.20	0.33	3.85	0.38	0.19	0.31	6.53	0.43	0.23	0.36
CogAgent-VQA	8.14	0.46	0.27	0.37	5.43	0.36	0.18	0.29	1.72	0.19	0.07	0.15	1.22	0.15	0.05	0.12
Intern-VLM-4khd	16.24	0.52	0.33	0.44	9.78	0.44	0.24	0.36	9.39	0.40	0.21	0.32	11.10	0.41	0.23	0.35
Interleaved Text-Image Baseline																
Gemini 1.5 Flash	27.13	0.57	0.36	0.52	19.24	0.48	0.25	0.43	21.88	0.48	0.27	0.42	23.51	0.56	0.33	0.51
GPT4o	32.41	0.67	0.47	0.56	24.56	0.62	0.39	0.51	22.08	0.56	0.33	0.46	26.74	0.63	0.41	0.49
Qwen-VL	3.81	0.19	0.07	0.16	3.45	0.20	0.06	0.17	4.20	0.17	0.05	0.05	1.48	0.11	0.03	0.10
Idefics-Mantis	7.71	0.38	0.21	0.35	7.67	0.22	0.12	0.20	9.36	0.34	0.17	0.31	3.49	0.33	0.17	0.31

Table 15: Detailed results on sampled subset of FeTaQA. BLEU, ROUGE-1,2,L are reported in the table. EQ - Explicit Questions, AQ - Answer-Mention Questions, IQ - Implicit Questions, VQ - Visual Questions. Best performing models are highlighted in red.

Prompt: You are given a table in which some entities in various table cells have been replaced by tokens of the type '{ENTITY-<entity_id>}. Each row of the table is in separate lines, and the columns are separated by '|'. Based upon the context of the table and using real-world knowledge, your task is to answer a question based upon the table by guessing the replaced entities of the table. You must perform this task in the following steps:

Step 1: Reason about what should be the answer to the question based upon the context of the table. The reasoning should be detailed and should be based upon the context of the table and the question, using real-world knowledge for answering the question and guessing various entities involved in finding the answer. IMPORTANT: You must explore any kind of reasoning -- numerical, logical, knowledge-based needed for answering the question.
Step 2: Based upon the reasoning provided, provide the answer to the question\n\nYou are given some question-answer samples to better format for providing the answer. IMPORTANT: You must give the answer in the format \"Step 2: <answer>\".

{few shot examples…}

Now, based upon the examples given above, you must understand the image given and follow the steps 1-2 to answer the question corresponding to the table represented in the image. It is IMPORTANT that you perform all the both the steps to the best possible extent to get the correct answer. You must follow the format of answers as demonstrated by the examples above. IMPORTANT: You must give the answer in the format 'Step 2:\n<answer>':

Table context: Table related to Team Chart in context of 2005 NASCAR Craftsman Truck Series.

Table:
Team | Truck(s) | # | Driver(s) | Primary Sponsor(s) | Listed Owner(s) | Crew Chief | \nBill Davis Racing | {ENTITY-1} | 5 | {ENTITY-2} | {ENTITY-1} | Bill Davis | Jeff Hensley | \nBill Davis Racing | {ENTITY-1} | 22 | {ENTITY-14} | Suntrust | Bill Davis | Doug Wolcott | \nBill Davis Racing | Toyota Tundra | 23 | {ENTITY-24} | Toyota Racing Development | Gail Davis | Greg Ely |...
...
Question: Name the trucks for scott neal
Step 1:

Figure 8: Prompt used for Partial Input Baseline

Prompt: You are given a table in the form of an image. In the table, some entities (mentioned in text form originally) have been replaced by images that represent them. Based upon the context of the table while using real-world knowledge, your task is to identify the entities corresponding to the images in the table and answer the question. You must perform this task in the following steps:

Step 1: Reason about what should be the answer to the question by identifying the relevant entities represented by images using the context of the table and the question. The reasoning should be detailed and should be based upon the context of the table and the question, using real-world knowledge for answering the question. IMPORTANT: You must explore any kind of reasoning -- numerical, logical, knowledge-based needed for disambiguating the entities and answering the question.
Step 2: Based upon the reasoning provided, provide the answer to the question.\n\nYou are also provided with some question-answer examples for better understanding the format of providing the answer:
{few shot examples…}
Now, based upon the examples given above, you must understand the image given and follow the steps 1-2 to answer the question corresponding to the table represented in the image. It is IMPORTANT that you perform all the both the steps to the best possible extent to get the correct answer. You must follow the format of answers as demonstrated by the examples above. IMPORTANT: You must give the answer in the format 'Step 2:
<answer>'.

Table context: Table related to Team Chart of 2005 NASCAR Craftsman Truck Series.\n\nTable:\nTeam | Truck(s) | # | Driver(s) | Primary Sponsor(s) | Listed Owner(s) | Crew Chief | \nBill Davis Racing | Toyota Tundra | 5 | Mike Skinner | Toyota Tundra | Bill Davis | Jeff Hensley | \nBill Davis Racing | Toyota Tundra | 22 | Bill Lester | Suntrust | Bill Davis | Doug Wolcott | \nBill Davis Racing | Toyota Tundra | 23 | Johnny Benson | Toyota Racing Development | Gail Davis | Greg Ely | \nBHR2 | Dodge Ram | 8 | Deborah Renshaw | Easy Care Service Contracts | Ray Montgomery | Bob Bissinger | \nBilly Ballew Motorsports | Chevrolet Silverado | 15 | Shane Hmiel | Kraft Foods | Billy Ballew | Richie Wauters | …
Question: Name the trucks for scott neal
Step 1:

Figure 9: Prompt used for Oracle entity replacement Baseline

Prompt: You are given certain infobox tables centred around a common topic. Each table lists its entries in separate lines, formatted as 'column_name: cell_content'. However, few entities in some of the tables have been replaced by images of those entities, denoted by tags of the format '{{ENTITY_IMAGE-<entity_id>}}' that represent those entities. Your task is to, based upon the context provided by the given table and the related prior tables, reason, predict and replace certain {{ENTITY_IMAGE}} tags by their original representative entity names and also provide their visual descriptions. Some example responses are given to better illustrate the task, and each new example starts with a line of '#'s. NOTE: YOU ARE ONLY PROVIDED IMAGES FOR THE MAIN TASK, NOT FOR THE EXAMPLE TASKS:
{few_shot_examples}
MAIN TASK:
Table context: {section_title} of {page_title}
{infobox_1}
{infobox_2}
{passage_info}
Now, based upon the examples above and the table(s) given below, your specific task is to replace the image tag(s) {entity_tags_string} mentioned in {col_header} column of the following {table_or_tables} with their original entities{same_entry_string}. In order to help you perform this task, you are also provided images corresponding to the tag(s) in the order {entity_tags_string}. You are also given relevant passages related ti the cell if there are any. Use them also. You must do this in the following steps:
Step 1: You must describe the relevant information about the entity that can be inferred from the given table context related to {section_title} of {page_title}. Ensure that this is as detailed as possible, and ONLY uses the information provided in the {num_tables} tables given above. Use your real-world knowledge to make as many inferences and form relationships as possible from the information provided in the image. IMPORTANT: It is of utmost importance that you DO NOT include information from the image in this description.
Step 2: You must visually describe the image(s) in complete detail, highlighting the important aspects based upon the context of the tables provided and the description obtained in Step 1. The visual descriptions must be based upon the fact that the image(s) occur in the entry {cell_text} representing a {col_header} in the context of {section_title} of {page_title}. Ensure that you consider other entries of {col_header} in other tables as well to make the visual description as accurate as possible. Using these visual descriptions, you must also identify the entities that are depicted using the different images in the context of the table.
Step 3: You must combine the information from Step 1 and Step 2 along with other attributes in the image, the context of the table(s) provided and real-world knowledge to provide the actual entry corresponding to {cell_text} of the {col_header} column from the table above.
Step 4: Based upon the response in Step 3, you must output the entities corresponding to the entity tag(s) present in the entry for Single in the format '{{ENTITY_IMAGE-<entity_id>}} -> <entity_string>'. It is VERY IMPORTANT that you follow this format while providing the output. You MUST list every entity tag in a…..

Figure 10: Prompt used for Image captioning Baseline

Prompt: You are given a table in the form of an image. In the table, some entities (mentioned in text form originally) have been replaced by images that represent them. Based upon the context of the table while using real-world knowledge, your task is to identify the entities corresponding to the images in the table and answer the question. You must perform this task in the following steps:

Step 1: Reason about what should be the answer to the question by identifying the relevant entities represented by images using the context of the table and the question. The reasoning should be detailed and should be based upon the context of the table and the question, using real-world knowledge for answering the question. IMPORTANT: You must explore any kind of reasoning -- numerical, logical, knowledge-based needed for disambiguating the entities and answering the question.
Step 2: Based upon the reasoning provided, provide the answer to the question.

You are also provided with some question-answer examples for better understanding the format of providing the answer:
{few-shot-examples}

Now, based upon the examples given above, you must understand the image given and follow the steps 1-2 to answer the question corresponding to the table represented in the image. It is IMPORTANT that you perform all the both the steps to the best possible extent to get the correct answer. You must follow the format of answers as demonstrated by the examples above. IMPORTANT: You must give the answer in the format 'Step 2:\n<answer>'.
Table context: {table_metadata}
Question: {question}

	Year	Title	Role	Notes
0	1995	Polio Water	Diane	Short film
1	1996	New York Crossing	Drummond	Television film
2	1997		Devon Stockard	
3	1999		Rocky	
4	1999		12-Year-Old Actress	

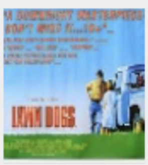
Figure 11: Prompt used for Table as an image baseline

Prompt: Answer in a sentence, using the table data given. The table consists of data in the form of text and images. Each row of the table has been represented using [] with data for each column in the row separated by a semi-colon.\n In the table, some entities (mentioned in text form originally) have been replaced by images that represent them. Based upon the context of the table while using real-world knowledge, your task is to identify the entities corresponding to the images in the table and answer the question. You must perform this task in the following steps:

Step 1: Reason about what should be the answer to the question by identifying the relevant entities represented by images using the context of the table and the question. The reasoning should be detailed and should be based upon the context of the table and the question, using real-world knowledge for answering the question. IMPORTANT: You must explore any kind of reasoning -- numerical, logical, knowledge-based needed for disambiguating the entities and answering the question.
Step 2: Based upon the reasoning provided, provide the answer to the question.

You are also provided with some question-answer examples for better understanding the format of providing the answer:
{few-shot-examples}

Now, based upon the examples given above, you must understand the image given and follow the steps 1-2 to answer the question corresponding to the table represented in the image. It is IMPORTANT that you perform all the both the steps to the best possible extent to get the correct answer. You must follow the format of answers as demonstrated by the examples above. IMPORTANT: You must give the answer in the format 'Step 2:\n<answer>'.
Table context: {table_metadata}
Question: {question}
Table: [[A, B, C]; [Suyash, <image1>, Red]; [Vivek, <image2>, Blue]]

<image1> -> 

, <image2> -> 

Figure 12: Prompt used for Interleaved Text-Image baseline