Multimodal Tabular Reasoning with Privileged Structured Information

Jun-Peng Jiang[1,2,3*] **Yu Xia**[3] **Hai-Long Sun**[1,2,3*] **Shiyin Lu**[3] **Qing-Guo Chen**[3]
Weihua Luo[3] **Kaifu Zhang**[3] **De-Chuan Zhan**[1,2] **Han-Jia Ye**[1,2†]
[1] School of Artificial Intelligence, Nanjing University, China
[2] National Key Laboratory for Novel Software Technology, Nanjing University, China
[3] AI Business, Alibaba Group.
{jiangjp, sunhl, zhandc, yehj}@lamda.nju.edu.cn
{daxiao.xy, running.lsy, qingguo.cqg, weihua.luowh, kaifu.zkf}@alibaba-inc.com

Abstract

Tabular reasoning requires complex, multi-step information extraction and logical inference, such as aggregation, comparison, or calculation over tabular data. While recent advances have leveraged large language models (LLMs) for reasoning over structured text tables, such high-quality textual representations are often unavailable in real-world settings, where tables typically appear as images. In this paper, we tackle the task of tabular reasoning directly from table images. Our core strategy is to leverage privileged structured information—specifically, the ground-truth structured table data available during training but inaccessible at test time—to enhance multimodal large language models (MLLMs). The key challenges lie in: accurately aligning visual representations with the structured information, particularly mapping the visual evidence to logical steps; and effectively transferring the reasoning skills learned during training to the MLLM for visual inference. To address these, we introduce TURBO (TabUlar Reasoning with Bridged infOrmation), a new framework for multimodal tabular reasoning using privileged information. TURBO benefits from a structure-aware reasoning trace generator based on DeepSeek-R1, which contributes to high-quality modality-bridged information. On this basis, TURBO repeatedly generates and selects advantageous reasoning traces, further enhancing the model's tabular reasoning ability. Experimental results demonstrate that, with limited (9k) data, TURBO achieves state-of-the-art performance ($+7.2\%$ vs. previous SOTA) across multiple datasets.

1 Introduction

Tabular understanding focuses on the extraction of information from various forms of tables, such as structured tables and table images [16, 31, 64, 73]. Going a step further, tabular reasoning [91, 77, 89] involves multi-step extraction, integration, and logical inference over tables to derive the final answer, demonstrating its potential in fields such as finance [95], healthcare [63], and scientific research [24].

With the advancement of large language models (LLMs) [4, 49, 8, 18, 75], particularly their increasing reasoning capabilities [48, 55, 21, 74], performing tabular reasoning tasks becomes feasible. Current efforts primarily focus on structured tables, where the input is typically the table's text or structured representations such as Markdown [85, 77, 89]. In contrast, in real-world scenarios, we can only have access to table images or screenshots rather than clean, structured tables [50, 90, 40]. Given the gap

[*] Work done during an internship at AI Business, Alibaba Group
[†] Corresponding author, email: yehj@lamda.nju.edu.cn.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

between training inputs and deployment conditions, a natural question arises: *can structured tables be leveraged as privileged information [72, 71] to improve the reasoning capabilities of MLLMs?*

Structured tables inherently contain rich information, allowing for fast and precise extraction of the content in each row and column [26, 76]. However, they often struggle to capture the full structural context, particularly in cases where the layout and relationships between rows and columns are complex [40]. On the other hand, while table images provide clear visual cues, current MLLMs still face challenges in effectively extracting and integrating corresponding information, particularly for complex tasks that require multi-step understanding [33, 80].

Beyond the gap between structured tables and images, while recent MLLMs have shown progress in domains such as mathematical reasoning [62, 66], their performance on tabular reasoning tasks remains limited [65, 92]. In our setting, the model is required to extract relevant content from table images, integrate information based on row and column relationships, and then perform logical inference or numerical computation to arrive at the correct answer. This multi-step process places high demands on both perception and reasoning, exposing clear limitations in current models.

Therefore, to enhance multimodal tabular reasoning capability with privileged structured tables, there are two main challenges. Given the complexity of structured tables, not all content is equally relevant to a given question. It is therefore crucial to identify and extract the most relevant information to guide model learning. Furthermore, once relevant content is identified, the next challenge is to effectively leverage it to improve the reasoning ability of MLLMs, enabling them to progressively construct accurate, multi-step reasoning paths.

To address the challenges, we propose TabUlar Reasoning with Bridged infOrmation (TURBO), a framework that integrates structured table information during training to enhance multimodal tabular reasoning. Our approach leverages modality-agnostic reasoning traces as bridged information, facilitating the knowledge transfer from LLMs to MLLMs. Additionally, we employ reinforcement learning techniques to further enhance the model's reasoning capabilities, allowing it to progressively improve its performance on complex tabular reasoning tasks.

Specifically, we construct a new pipeline tailored for tabular reasoning. We start by feeding structured markdown tables and their corresponding QA pairs into DeepSeek-R1 [21], the SOTA reasoning LLM, to generate structure-aware chains of thought to bridge structured tables and images. This process yields high-quality reasoning data in the form of [question, reasoning, answer] triples, which effectively filter out redundant content and emphasize question-relevant reasoning traces. We then conduct supervised fine-tuning (SFT), equipping the MLLM with initial tabular reasoning capabilities. Building on this foundation, by iteratively sampling and selecting the relative advantages in reasoning paths as GRPO method [61], we further enhance the reasoning capability, progressively strengthening the reasoning performance. Experimental results demonstrate that our TURBO achieves significant improvements ($+7.2\%$ over baselines) with limited data, showing high generalization ability and interpretability. Our contributions are threefold:

- We highlight the practical value of using structured tables as privileged information during training to enhance reasoning over table images, as such structured tables are often unavailable at inference time in real-world scenarios.
- We propose a new framework that utilizes privileged structured information to bridge the modality gap and enhance the tabular reasoning capabilities of MLLMs, combining structure-aware reasoning trace generation with iterative optimization.
- Experimental results show that our method achieves substantial performance improvements over datasets with limited training data. Furthermore, the resulting model demonstrates strong robustness and interpretability in its reasoning process.

2 Related Work

2.1 Tabular Reasoning

Tabular data is of great importance due to its broad applicability in real-world domains, encompassing a variety of learning tasks from standard classification and regression tasks [82, 38, 84, 28] to open-world challenges [29, 6, 30, 27, 39, 83, 5]. Besides, tabular reasoning is a part of tabular understanding, which involves comprehending the information contained within a table and can be broken down into several tasks, such as Table Structure Recognition (TSR) [59, 58], Table Detection

(TD) [19, 35], and Table Question Answering (TQA) [9, 69, 31]. Traditional methods, whether OCR-based [2, 13, 20] or OCR-free [46, 32, 17, 73, 90], have made significant strides in recognizing the structure and content of tables. However, we focus on more challenging tabular reasoning tasks.

Tabular reasoning involves multi-step information extraction and logical inference over tabular data. In the field of tabular reasoning, several recent works have made notable progress. OPENRT [91] is a open-source framework designed for reasoning over tabular data. It enables the reproduction of existing table pre-training models for fair performance comparison and supports rapid development of new models. [15] explores different table representations and directly prompts LLMs and MLLMs, investigating how structural formats influence reasoning capabilities. Chain-of-Table [77] employs in-context learning to iteratively generate operations and update the table, allowing LLMs to construct reasoning chains where each step builds upon the previous output. HIPPO [40] represents tables using both textual and visual modalities, and optimizes MLLMs to learn richer and more comprehensive information with DPO [57]. However, these works primarily focus on reasoning over structured tables and achieve only limited performance. Since structured tables are rarely available in real-world settings, reasoning over table images becomes a more practical choice.

2.2 MLLMs for Reasoning

The field of multimodal large language models (MLLMs) has advanced significantly, particularly in integrating visual and textual processing. Modern MLLMs combine visual encoders [56, 68, 87, 25, 67], LLMs, and fusion modules. Models like BLIP-2 [34] and InstructBLIP [14] use a Q-Former to bridge vision encoders and frozen LLMs. MiniGPT-4 [93] combines Q-Former with a linear projector for better alignment between vision and LLMs. LLaVA [37] employs an MLP projector to improve vision-LLM alignment. Alongside these open-source advancements, proprietary models like GPT-4o/o1 [47, 1], Gemini2.5pro [70], and Qwen2.5-VL-Plus/MAX [3] have excelled in benchmarks and real-world applications.

Recent developments in MLLMs have led to significant improvements in reasoning tasks involving both textual and multimodal data [48, 21, 54, 66, 53, 43]. Most current approaches rely on Chain-of-Thought (CoT) techniques [78] to train MLLMs for sequential reasoning. Notable data-driven initiatives include Math-LLaVA [62], which introduced the MathV360K dataset, and MAmmoTH-VL [22], which builds large multimodal CoT datasets at scale. Another research direction focuses on improving vision-text alignment. MAVIS [88] fine-tunes a math-specific vision encoder with a curated set of captions, while Math-PUMA [96] aligns modalities by utilizing the KL divergence in next-token prediction distributions. In our paper, we focus on reasoning over table images, where structural information is not explicitly available and must be inferred through visual cues and multi-step reasoning to reach the correct answer.

3 Preliminaries

In this section, we first introduce the notation of multimodal tabular reasoning with privileged structured information task, followed by some preliminary experiments and analyses.

3.1 Multimodal Tabular Reasoning with Privileged Structured Information

Tabular reasoning refers to the task of answering questions that require understanding and logical inference over tabular data. This often involves multi-step reasoning, including locating relevant cells, integrating information across rows and columns, and performing operations such as comparison, aggregation, or arithmetic.

In real-world applications, high-quality structured tables T, such as markdown tables, are often difficult to obtain. Instead, what is more readily available are table images V, such as screenshots or scanned documents. These image-based tables present significant challenges for reasoning, as the model must perceive both the content and the visual layout before performing any logical inference.

To address this, our goal is to enhance multimodal large language models $\mathcal{M}$ to generate a response $X_{\mathtt{a}}$ for the tabular reasoning question Q over table images V by leveraging structured tables T as privileged information during training. While structured tables are *not accessible at inference time*, they offer rich semantics and precise structure that can guide the model *during training*. By

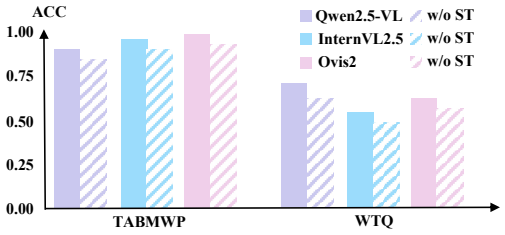

Figure 1: Impact of structured tables (ST). We evaluate the effect of providing structured tables or not as input across Qwen2.5-VL [3], InternVL2.5 [11], and Ovis2 [42] on the TABMWP [41] and WTQ [52] datasets. The results show a consistent performance drop when structured tables are removed, highlighting their importance for effective tabular reasoning.

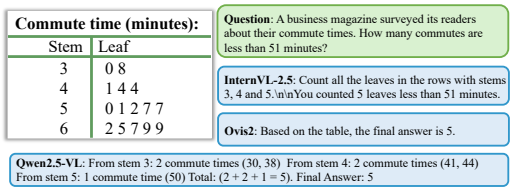

Figure 2: Challenge of current MLLMs on tabular reasoning. This example illustrates the limitations of current MLLMs in performing multi-step reasoning and numerical computation. The task requires sequential reasoning based on the image, but existing models often lack such capabilities or make errors during intermediate steps, ultimately leading to incorrect answers.

combining visual table inputs with structured supervision, we aim to improve the MLLM's ability to perform accurate and interpretable reasoning over table images. Our goal is for the final answer, $X_{\mathtt{a}}$, to be as consistent as possible with the correct answer $Y_{\mathtt{a}}$.

$$\min_{\mathcal{M}} \sum_{i=1}^{N} \ell(\mathcal{M}(V, Q|T), Y_{\mathtt{a}}) \ . \tag{1}$$

ℓ is the loss function that measures the discrepancy between prediction and ground-truth. "|" indicates that the answer is generated conditioned on the corresponding structured table T. Our objective is to enhance the $\mathcal{M}$'s reasoning capability with the privileged structured information. In situations where structured tables is not available in real-world scenarios, we expect $\mathcal{M}$ can still provide accurate answers given only the table image V. Therefore, during the inference phase, we measure the performance of $\mathcal{M}$ based on its prediction accuracy given any test image.

3.2 Analysis on MLLMs in Tabular Reasoning

To conduct our preliminary study, we select several of the most advanced and widely adopted open-source MLLMs, including Qwen2.5-VL [3], InternVL2.5 [11], and Ovis2 [42], and evaluate them on TABMWP [41] and WikiTableQuestions (WTQ) [52]. We randomly sample 100 examples from each dataset for testing. Since tabular reasoning tasks involve open-ended question answering, it is often challenging to directly judge the correctness of model outputs. We employ Qwen2.5-72B-Instruct [79] as an external evaluator. Specifically, we provide the LLM with the question, ground-truth answer, and the model's prediction. By converting the evaluation into a single-modality setting, this approach reduces ambiguity and improves the reliability of the evaluation, offering a more accurate assessment of MLLMs' reasoning capabilities.

To investigate the role of structured tables in tabular reasoning, we conduct preliminary experiments using MLLMs. Specifically, we compare the model's performance when reasoning is performed with access to structured tables versus relying solely on table images during inference.

Structured Information Matters in Tabular Reasoning. Figure 1 presents a comparison between inference results with and without access to structured tables. We observe a significant drop in performance across all models when structured tables are not provided. This highlights the difficulty MLLMs face when relying solely on table images, as they must implicitly recover the table's structure and relevant relationships before performing reasoning. In contrast, structured tables offer clean and explicit content, which greatly facilitates accurate information extraction. These results underscore the importance of leveraging structured representations to guide model reasoning, even if they are only available during training.

MLLMs Lack Explicit Reasoning Traces. Both WTQ and TABMWP contain complex, reasoning-intensive questions that require multi-step inference and arithmetic operations to reach the correct answer, as in the case in Figure 2. However, our analysis reveals that current MLLMs often fall

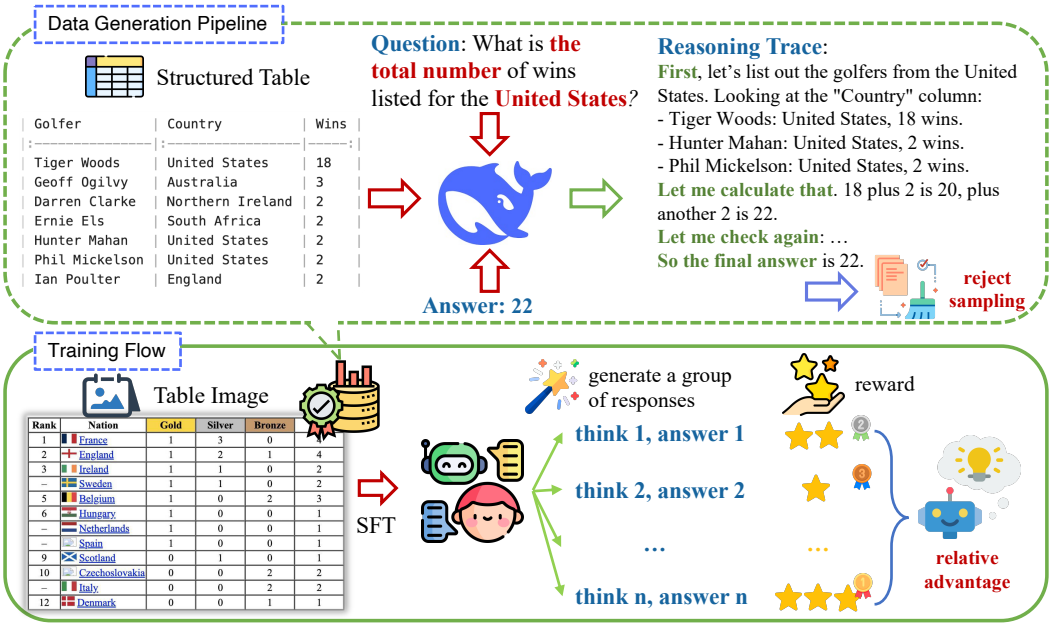

Figure 3: Data generation pipline and training flow of our TURBO framework. We first input the structured table, question, and answer into DeepSeek to generate a reasoning trace, and apply reject sampling to improve data quality. Then, supervised fine-tuning (SFT) is performed on the collected data for privileged information alignment. Finally, we generate a group of answers for each question and compute their relative advantages based on their reward to further reinforce the reasoning ability.

short in these scenarios. Despite having access to table images, these models tend to produce final answers directly, frequently bypassing any clear intermediate reasoning steps. In the absence of explicit prompting for chain-of-thought generation, their outputs rarely exhibit clear reasoning traces. This lack of step-by-step thinking not only undermines interpretability but also frequently results in incorrect predictions. Furthermore, even Qwen shows potential in reasoning traces, there are critical intermediate mistakes, leading to the final wrong answer. These observations highlight the limitations of current MLLMs in tabular reasoning and underscore the need for methods that can explicitly leverage structured reasoning capability and align it with visual representations.

4 Multimodal Tabular Reasoning with Bridged Information

In this section, we introduce our TURBO framework (as shown in Figure 3) of multimodal tabular reasoning with privileged structured information through a structure-aware reasoning trace generator. Then we sample and select the advantageous reasoning paths, further enhancing the reasoning ability.

4.1 Bridging Structured Tables and Images

A central challenge in multimodal tabular reasoning arises from the intrinsic modality gap between structured tables and table images. These two modalities, though semantically aligned, are processed very differently by MLLMs, leading to differences in their reasoning behavior.

Structured tables, such as those in Markdown or HTML, explicitly encode row-column relationships, enabling accurate and fine-grained access to cell contents [7, 52]. This makes them particularly suitable for tasks that require semantic-level reasoning. However, they often fail to capture complex layout nuances such as merged cells, multi-level headers, or visual hierarchies, which are common in real-world tables [15, 40]. In contrast, table images naturally preserve these structural cues through visual patterns like borders, spacing, and alignment. Yet, due to OCR errors, visual noise, or resolution limits, accurately extracting textual content from table images remains a significant challenge [44]. This inherent modality gap highlights the complementary nature of structured tables and table images—each captures essential but distinct aspects of tabular information.

Unified Reasoning across Modalities. However, despite their representational differences, both structured tables and table images ultimately convey the same underlying information relevant to reasoning tasks. From a reasoning perspective, the inference trajectory—*i.e.*, the logical path leading from question to answer—would remain consistent across both modalities. A correct reasoning chain, once established, can guide models to identify the relevant content and apply appropriate reasoning steps, regardless of whether the input is a structured table or its visual counterpart. This highlights that the essence of tabular reasoning lies not in the input format but in the ability to extract and compose the necessary information into a coherent, accurate reasoning process.

Therefore, it is natural to leverage the strong reasoning capabilities of current large language models to extract high-quality reasoning chains for tabular data. These structured chains of thought are modality-independent and can serve as valuable supervision signals for improving multimodal tabular reasoning. To generate structure-aware reasoning traces, we leverage the available structured tables during training. Specifically, for each markdown-formatted table T and its associated question-answer pair (Q, A), we input (T, Q, A) into DeepSeek-R1 [21], a state-of-the-art reasoning LLM. By providing the ground-truth answer A alongside the question and table, we guide the model to produce a coherent and accurate reasoning process $R = \{r_1, r_2, ..., r_n\}$ that logically connects the question to the correct answer.

Reject Sampling over Reasoning. This approach produces high-quality supervision triplets (Q, R, A), where the reasoning trace R captures the key logical steps, relevant table elements, and necessary operations to arrive at the correct answer. To further improve data quality, we apply additional reject sampling: overly verbose or contradictory reasoning traces—often caused by inconsistencies between DeepSeek's generated answer and the provided ground-truth answer—are removed. We also employ an auxiliary strong LLM (*e.g.*, Qwen2.5-72B-Instruct) to assess the coherence and correctness of each trace, ensuring that only faithful and focused reasoning chains are retained. After reject sampling, we obtain approximately 9k high-quality examples that serve as reliable supervision for enhancing multimodal tabular reasoning.

In summary, we introduce the bridged information that successfully aligns structured tables with table images through reasoning chains. Our structure-aware reasoning trace generator provides high-quality reasoning traces, facilitating the transfer of reasoning from structured inputs to images.

4.2 Enhancing MLLMs with Reasoning Capability

Preliminary Exploration of Reasoning Capability. The most straightforward way to enhance MLLMs with reasoning ability is to use supervised fine-tuning (SFT) on our high-quality reasoning data. We construct the following system prompt to guide the model's learning process: `You are a helpful assistant. For each question, first think through your reasoning, then provide an answer. Format your response as: <think>Your reasoning process</think><answer>Your final answer</answer>`. Each training ground truth is formatted as follows: `<think>`R`</think><answer>`A`</answer>`. Through SFT, this process enables MLLMs to learn the "reasoning-first, answer-later" paradigm. Moreover, since the reasoning traces are modality-invariant, the reasoning ability can be preliminarily transferred from LLMs to MLLMs across modalities.

Reinforcing Multimodal Reasoning Capability in MLLMs. Inspired by [21], we adopt Group Relative Policy Optimization (GRPO) as the reinforcement learning algorithm to further enhance the reasoning capability. Unlike SFT, which minimizes token-level prediction errors, GRPO leverages policy gradients derived from reward signals. This enables the model to explore a broader range of possible solutions, encouraging more diverse and complex reasoning behaviors [21, 36].

Specifically, we define questions as Q and the current policy as $\pi_{\theta_{old}}$. For each question $q \in Q$, a group of responses $\{o_1, o_2, ..., o_G\}$ is generated from $\pi_{\theta_{old}}$. A fixed reference policy π_{ref} is also used to regularize training. The optimization objective for the updated policy π_θ is defined as:

$$J(\theta) = \mathbb{E}_{q\sim Q,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}} \left[\frac{1}{G}\sum_{i=1}^{G} \min\left(\frac{\pi_\theta\left(o_i|q\right)}{\pi_{\theta_{\text{old}}}\left(o_i|q\right)} A_i, \text{clip}\left(\frac{\pi_\theta\left(o_i|q\right)}{\pi_{\theta_{\text{old}}}\left(o_i|q\right)}, 1-\epsilon, 1+\epsilon \right) A_i \right) - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) \right] \quad (2)$$

Here, ϵ is the clipping threshold, and β is the weight for the KL divergence penalty. The advantage A_i for each response o_i is computed relative to the group as:

$$A_i = \frac{r_i - \text{mean}\left(\{r_1, r_2, \ldots, r_G\}\right)}{\text{std}\left(\{r_1, r_2, \ldots, r_G\}\right)}. \tag{3}$$

This approach normalizes the reward across the group to obtain a relative advantage signal. The KL divergence term is defined as:

$$D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right) = \frac{\pi_{\text{ref}}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)} - \log\left(\frac{\pi_{\text{ref}}\left(o_i \mid q\right)}{\pi_\theta\left(o_i \mid q\right)}\right) - 1. \tag{4}$$

By leveraging relative advantage, the model is able to perform self-comparison across its own set of generated responses, identifying which responses or reasoning paths are relatively more effective. Meanwhile, the KL divergence term acts as a regularizer, ensuring that the updated policy does not deviate excessively from the reference model in a single optimization step. This constraint helps maintain training stability by preventing abrupt policy shifts, enabling the model to improve progressively in a controlled and stable manner.

Unlike standard PPO [60], GRPO removes the need for a critic network by estimating the relative advantage within each sampled group. This not only simplifies implementation but also significantly reduces computational overhead. In our tabular reasoning setting, the reward function is composed of two parts: format reward and accuracy reward.

The format reward encourages the model to adhere to a structured reasoning format by using `<think></think>` and `<answer></answer>` tags in its response. This design promotes a two-step process where the model first articulates a chain of reasoning based on its understanding of the table, and then explicitly states the final answer derived from that reasoning. By rewarding outputs that follow this format, we guide the model to develop more interpretable and systematic reasoning behavior. The accuracy reward is rule-based and evaluates whether the content in the `<answer></answer>` tags matches the ground truth labels. This component directly incentivizes correctness in the final answer, ensuring that the reasoning process ultimately leads to accurate outcomes.

In summary, reward-guided GRPO efficiently enhances the model's reasoning ability in tabular tasks by selecting superior responses within a group, enabling self-improvement without the overhead of a value function. Through this two-stage learning framework, our model develops a strong tabular reasoning capability, effectively bridging structured and unstructured modalities while enabling robust, interpretable reasoning across table images.

5 Experiments

In this section, we first present the experimental setup, including datasets, baselines, and implementation details. We then demonstrate the effectiveness of our approach through comprehensive experiments and ablation studies. Finally, case studies further highlight the interpretability and robustness of our method in complex tabular reasoning scenarios.

5.1 Experimental Setup

Implementation Details. We conduct all experiments based on the Ovis2 [42]. Our training pipeline consists of two stages: supervised fine-tuning (SFT) and reinforcement learning (RL). During the SFT stage, we train our model using 4×A100 GPUs for 1 hour with a batch size of 128 and a learning rate of 2e-6. This stage allows the model to learn structure-aware reasoning patterns from high-quality [question, reasoning, answer] triples. In the RL stage, we further enhance the model's reasoning ability by applying GRPO with 8×A100 GPUs for 24 hours. We use a smaller learning rate of 5e-7 with the same batch size of 128. For each question, the model generates 16 candidate responses, which are then used to compute relative advantages and guide the optimization process.

Evaluation Benchmarks. Following HIPPO [40], we evaluate the tabular reasoning capability of our model on a diverse set of publicly available benchmarks. Specifically, for Tabular Question Answering (TQA) tasks, we use five representative datasets: TABMWP [41], WTQ [52], HiTab [12], and TAT-QA [94]. We export FeTaQA [45] due to its evaluation metric being BLEU [51], which

Table 1: Accuracy comparison across tabular reasoning benchmarks. Our model, TURBO, consistently achieves the best overall performance, demonstrating strong reasoning ability and effective integration of privileged structured table information. Even when compared to the strongest baseline Qwen2.5-VL, TURBO achieves an average improvement of 7.2%.

Method	Question Answering				Fact Verification		MMMU	Average
	TABMWP	WTQ	HiTab	TAT-QA	TabFact	InfoTabs		
InternVL-2.5 [11]	90.88	43.19	45.94	34.97	66.46	55.50	41.46	54.06
Qwen2.5-VL [3]	92.48	65.85	67.09	70.54	83.01	77.91	42.07	71.28
MiniCPM-V-2.6 [81]	83.68	47.97	56.53	51.55	78.48	73.03	31.10	60.33
HIPPO [40]	87.34	55.71	63.13	61.40	82.29	75.70	-	-
HIPPO w/o ST	85.83	49.10	57.23	56.22	80.20	72.74	35.98	62.47
Table-LLaVA [92]	53.20	16.62	7.87	10.49	57.62	66.78	17.68	32.89
TabPedia [90]	10.66	23.53	6.54	13.08	35.49	2.43	2.44	13.45
Ovis2 [42]	92.00	58.76	68.59	47.67	80.80	74.11	48.17	67.16
Ovis2-CoT	92.12	60.80	66.43	48.70	81.61	72.46	50.61	67.53
TURBO	**96.75**	**67.80**	**72.15**	**73.21**	**85.81**	**81.89**	**57.32**	**76.42**

does not focus on the accuracy of question answering. For Table Fact Verification (TFV) tasks, we include TabFact [10] and InfoTabs [23]. To further assess the robustness of our approach, we conduct additional experiments on MMMU [86], evaluating our model on all table-related questions within this challenging multimodal benchmark. This allows us to verify whether the reasoning capabilities learned through our framework generalize to unseen and diverse table images in complex real-world settings. For TQA, we evaluate model performance using Accuracy (ACC), and in TFV, we use the binary classification accuracy for TabFact (true/false outputs) and multi-class accuracy for InfoTabs (entail/contradict/neutral outputs). We employ Qwen2.5-72B-Instruct [79] as an external evaluator. Specifically, we provide the LLM with the question, ground-truth answer, and the model's prediction.

Comparison Methods. For comprehensive comparisons, we select leading open-source MLLMs and those in tabular understanding, including Qwen2.5-VL-7B [3], InternVL-2.5-8B [11], MiniCPM-V-2.6-8B [81], Ovis2-8B [42], Table-LLaVA-7B [92], TabPedia-7B [90], and HIPPO-8B [40].

5.2 Main Results

In this experiment, we evaluate the table reasoning effectiveness of our TURBO and baseline models on both the TQA, TFV tasks, and MMMU. For the HIPPO method, which leverages both structured tables and table images during training and inference, we report its performance under the original setting with structured tables and table images as input. Additionally, we include HIPPO without structured tables as "HIPPO w/o ST", a variant that takes only table images as input, which better aligns with our setting and serves as a fairer comparison for evaluating purely from table images.

To ensure a fair comparison with HIPPO, we follow the same data sampling strategy by randomly selecting 10k examples in these training datasets as our training set, which do not overlap with our test set. Furthermore, to evaluate the generalization capabilities of TURBO on completely unseen data distributions and formats, we included the challenging MMMU benchmark as an additional test set. We then apply our structure-aware reasoning trace generator to produce reasoning traces. To maintain high data quality, we further conduct reject sampling on incorrect or low-quality generations. As a result, we retain approximately 9k high-quality training samples for SFT and RL for 1 epoch. Results are in Table 1.

Across all datasets, our model TURBO achieves the best overall performance, demonstrating clear and consistent improvements over existing baselines. Specifically, TURBO surpasses the strong tabular understanding baseline HIPPO up to a relative improvement of 21.7% over TQA datasets (in WTQ), and by 8.2% over TFV datasets (in InfoTabs). Notably, compared to the variant HIPPO "w/o ST" (which only uses table images and is thus more comparable to our real-world setting), on average, our method achieves a 26.8% improvement on TQA and 9.8% on TFV, highlighting the effectiveness of our reasoning supervision and privileged structured information.

Furthermore, on the challenging MMMU benchmark, which evaluates table reasoning in real-world multimodal scenarios, TURBO reaches 57.32, outperforming the previous best method Ovis2-CoT

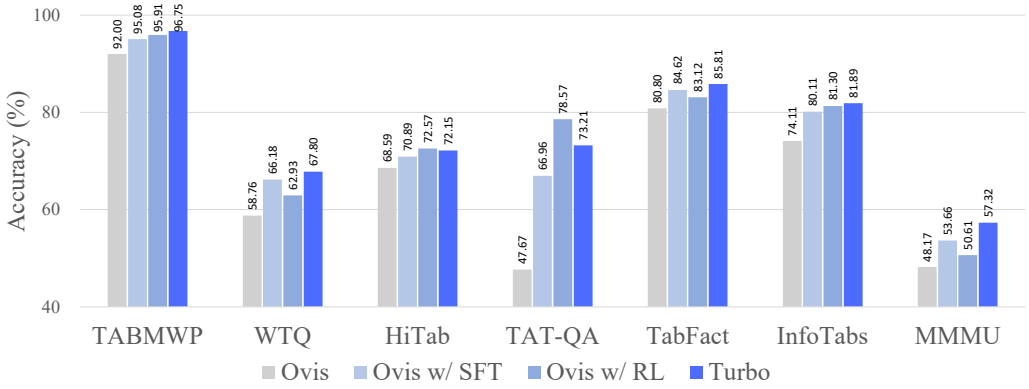

Figure 4: Ablation study of components in TURBO on each dataset. TURBO achieves consistent and significant improvements over the Ovis baseline on all datasets, demonstrating the effectiveness of each introduced component.

Year	Project A	Project B
0	-$20,000	-$24,000
1	13,200	14,100
2	8,300	9,800
3	3,200	7,600

Question: Maxwell Software, Inc., has the following mutually exclusive projects.Suppose the company uses the NPV rule to rank these two projects. Which project should be chosen if the appropriate discount rate is 15 percent?
A: Project A
B: Project B

Turbo: <think>Okay, let's see. The problem is about choosing between Project A and Project B using the NPV rule **with a 15% discount rate**.
First, I need to calculate the NPV for each project. **Looking at the table**, the initial investment for Project A is **-$20,000,** and for Project B, it's **-$24,000**. The cash flows for each year. The NPV formula is … discounted by the discount rate (15% or 0.15).
For Project A:\nYear 0: -$20,000 (initial investment)\nYear 1: **$13,200 / (1+0.15)^1 = $13,200 / 1.15 ≈ $11,478.26** … NPV_A = -$20,000 + $11,478.26 + $6,279.07 + $2,104.53 ≈ -$138.14
For Project B:\nYear 0: -$24,000 (initial investment)\nYear 1: **$14,100 / (1+0.15)^1 = $14,100 / 1.15 ≈ $12,260.87** … NPV_B = -$24,000 + $12,260.87 + $7,401.71 + $4,997.45 ≈ $659.31
Comparing the NPVs, Project B has a higher NPV than Project A. Therefore, Project B should be chosen.</think>
<answer>B</answer>

Figure 5: Case study of Turbo. Our method accurately extracts relevant information from the table image (highlighted in blue), formulates the appropriate calculation expressions (in pink), and performs step-by-step reasoning to arrive at the correct final answer (in green). This demonstrates the effectiveness of TURBO in both visual grounding, mathematics, and logical reasoning.

by 13%, and surpassing HIPPO by 59.3% relative improvement, showcasing the robustness and generalization of our model under diverse and complex table formats.

These results validate that our two-stage learning framework not only equips MLLMs with strong reasoning capability, but also highlights its effectiveness in bridging the gap between structured tables and visual table inputs, enabling the model to absorb and utilize privileged structural information.

5.3 Further Studies

Ablation Study. To better understand the contribution of each component in our two-stage framework, we conducted an ablation study on our baseline model. As shown in Figure 4, we compare the following variants: Baseline: The base model Ovis2 without reasoning supervision. "w/ SFT": Supervised Fine-tuned with our structure-aware reasoning traces. "w/ RL": Optimized with reward-driven reinforcement learning stage. TURBO: The full model trained with both SFT and RL stages.

From the results, we observe consistent improvements across all evaluation tasks. Compared to the baseline, adding SFT alone brings 10.1% relative improvement on average, demonstrating that structure-aware reasoning traces help the model better understand table semantics. The RL stage also introduces 11.8% performance gains, particularly on complex QA datasets like TAT-QA, showing that relative advantage can refine reasoning fidelity and answer consistency. Finally, the full TURBO model achieves the best performance across the board, 13.8% on average. This confirms that both stages in our framework are complementary and jointly crucial for maximizing reasoning ability.

Case Studies. To further demonstrate the effectiveness of our approach, we present a case in Figure 5 that showcases the reasoning capability of our TURBO. In this example, the task involves interpreting a table image and answering a complex question that requires multiple-step reasoning, including visual grounding, mathematics, and logical reasoning. More cases and experiment results, such as more ablation studies about the extension of our framework, more experiments on inference time, can be found in Appendix B.

In the baseline model, answers are generated directly without an interpretable reasoning process. After the SFT stage, the model begins to exhibit basic reasoning behaviors but still makes shortcuts, such as incorrect calculations. With RL tuning, TURBO chooses advantageous answers in a group of responses, generates step-by-step reasoning, accurately locates key table elements, and produces interpretable answers, demonstrating clear improvements in both reasoning depth and accuracy.

6 Conclusion

In this paper, we propose TURBO, a new multimodal large language model tailored for real-world tabular reasoning tasks, where structured tables are used as privileged information. Our method leverages a two-stage training paradigm that integrates structured reasoning supervision into MLLMs. Specifically, we first introduce a high-quality, structure-aware reasoning trace generation mechanism, enabling the model to internalize a "think-then-answer" paradigm during supervised fine-tuning. We then further refine the reasoning capabilities via reinforcement learning, aligning the model's outputs with more accurate and interpretable reasoning behaviors. Our key contributions to novelty are: 1) a novel application of learning with privileged information; 2) the modality-bridging data heneration pipeline; 3) a synergistic and effective training framework. Extensive experiments show that TURBO consistently achieves state-of-the-art performance among open-source MLLMs. We hope this work inspires future research into multimodal reasoning, especially in developing more systematic and interpretable frameworks for real-world tasks.

Acknowledgments and Disclosure of Funding

This work is partially supported by National Key R&D Program of China (2024YFE0202800), NSFC (62376118), Collaborative Innovation Center of Novel Software Technology and Industrialization, Postgraduate "AI+" Research and Practical Innovation Project (KYCX25_0328).

References

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

[2] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003, 2021.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Hao-Run Cai and Han-Jia Ye. Feature-aware modulation for learning from temporal tabular data. In *Advances in Neural Information Processing Systems 39*, 2025.

[6] Hao-Run Cai and Han-Jia Ye. Understanding the limits of deep tabular methods with temporal shift. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 6366–6386, 2025.

[7] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C Hsieh, Deborah A Wallach, Mike Burrows, Tushar Chandra, Andrew Fikes, and Robert E Gruber. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems (TOCS)*, 26(2):1–26, 2008.

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45, 2024.

[9] Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*, 2020.

[10] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[12] Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. *arXiv preprint arXiv:2108.06712*, 2021.

[13] Cheng Da, Peng Wang, and Cong Yao. Multi-granularity prediction with learnable fusion for scene text recognition. *arXiv preprint arXiv:2307.13244*, 2023.

[14] Wenliang Dai, Junnan Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*, 2, 2023.

[15] Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. Tables as texts or images: Evaluating the table reasoning ability of llms and mllms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 407–426, 2024.

[16] Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: Table understanding through representation learning. *ACM Special Interest Group on Management of Data Record*, 51(1):33–40, 2022.

[17] Hao Feng, Zijian Wang, Jingqun Tang, Jinghui Lu, Wengang Zhou, Houqiang Li, and Can Huang. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023.

[18] Jiale Fu, Yuchu Jiang, Junkai Chen, Jiaming Fan, Xin Geng, and Xu Yang. Speculative ensemble: Fast large language model ensemble via speculation. *arXiv preprint arXiv:2502.01662*, 2025.

[19] Azka Gilani, Shah Rukh Qasim, Imran Malik, and Faisal Shafait. Table detection using deep learning. In *2017 14th IAPR international conference on document analysis and recognition*, volume 1, pages 771–776. IEEE, 2017.

[20] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. Xylayoutlm: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4583–4592, 2022.

[21] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[22] Jarvis Guo, Tuney Zheng, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Yizhi Li, Graham Neubig, Wenhu Chen, and Xiang Yue. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. *arXiv preprint arXiv:2412.05237*, 2024.

[23] Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. Infotabs: Inference on tables as semi-structured data. *arXiv preprint arXiv:2005.06117*, 2020.

[24] Vivek Gupta, Shuo Zhang, Alakananda Vempala, Yujie He, Temma Choji, and Vivek Srikumar. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3268–3283, 2022.

[25] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[26] Jonathan Herzig, Paweł Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Martin Eisenschlos. Tapas: Weakly supervised table parsing via pre-training. *arXiv preprint arXiv:2004.02349*, 2020.

[27] Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

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

[29] Jun-Peng Jiang, Han-Jia Ye, Leye Wang, Yang Yang, Yuan Jiang, and De-Chuan Zhan. Tabular insights, visual impacts: transferring expertise from tables to images. In *Proceedings of the 41st*

International Conference on Machine Learning, pages 21988–22009, 2024.

[30] Jun-Peng Jiang, Tao Zhou, De-Chuan Zhan, and Han-Jia Ye. Compositional condition question answering in tabular understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 27831–27850, 2025.

[31] Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. A survey on table question answering: recent advances. In *China Conference on Knowledge Graph and Semantic Computing*, pages 174–186. Springer, 2022.

[32] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Donut: Document understanding transformer without ocr. *arXiv preprint arXiv:2111.15664*, 7(15):2, 2021.

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

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[35] Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, 2020.

[36] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, and Kaipeng Zhang. Think or not think: A study of explicit thinking in rule-based visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.16188*, 2025.

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[38] Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and Han-Jia Ye. Talent: A tabular analytics and learning toolbox. *arXiv preprint arXiv:2407.04057*, 2024.

[39] Si-Yang Liu and Han-Jia Ye. Tabpfn unleashed: A scalable and effective solution to tabular classification problems. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 40043–40068, 2025.

[40] Zhenghao Liu, Haolan Wang, Xinze Li, Qiushi Xiong, Xiaocui Yang, Yu Gu, Yukun Yan, Qi Shi, Fangfang Li, Ge Yu, et al. Hippo: Enhancing the table understanding capability of large language models through hybrid-modal preference optimization. *arXiv preprint arXiv:2502.17315*, 2025.

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

[42] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

[43] Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, et al. Ovis2. 5 technical report. *arXiv preprint arXiv:2508.11737*, 2025.

[44] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition*, pages 947–952. IEEE, 2019.

[45] Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022.

[46] Ahmed Nassar, Nikolaos Livathinos, Maksym Lysak, and Peter Staar. Tableformer: Table structure understanding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4614–4623, 2022.

[47] OpenAI. Gpt-4o: Hello gpt-4o. Technical report, OpenAI, 2024.

[48] OpenAI. Learning to reason with llms. Technical report, OpenAI, 2024.

[49] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[50] Shubham Singh Paliwal, D Vishwanath, Rohit Rahul, Monika Sharma, and Lovekesh Vig. Tablenet: Deep learning model for end-to-end table detection and tabular data extraction from scanned document images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 128–133, 2019.

[51] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
[52] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.
[53] Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.
[54] QwenTeam. Qvq: To see the world with wisdom. Technical report, QwenTeam, Alibaba, 2024.
[55] QwenTeam. Qwq: Reflect deeply on the boundaries of the unknown. Technical report, QwenTeam, Alibaba, 2024.
[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
[57] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.
[58] Mahmoud Salaheldin Kasem, Abdelrahman Abdallah, Alexander Berendeyev, Ebrahem Elkady, Mohamed Mahmoud, Mahmoud Abdalla, Mohamed Hamada, Sebastiano Vascon, Daniyar Nurseitov, and Islam Taj-Eddin. Deep learning for table detection and structure recognition: A survey. *ACM Computing Surveys*, 56(12):1–41, 2024.
[59] Sebastian Schreiber, Stefan Agne, Ivo Wolf, Andreas Dengel, and Sheraz Ahmed. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition*, volume 1, pages 1162–1167. IEEE, 2017.
[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
[61] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.
[62] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024.
[63] Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce Ho, Carl Yang, and May D Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. *arXiv preprint arXiv:2401.07128*, 2024.
[64] Alexey Shigarov. Table understanding: Problem overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13(1):e1482, 2023.
[65] Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, et al. Tablegpt2: A large multimodal model with tabular data integration. *arXiv preprint arXiv:2411.02059*, 2024.
[66] Hai-Long Sun, Zhun Sun, Houwen Peng, and Han-Jia Ye. Mitigating visual forgetting via take-along visual conditioning for multi-modal long cot reasoning. In *ACL*, 2025.
[67] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, et al. Parrot: Multilingual visual instruction tuning. In *International conference on machine learning*, 2025.
[68] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.
[69] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.
[70] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
[71] Vladimir Vapnik, Rauf Izmailov, et al. Learning using privileged information: similarity control and knowledge transfer. *J. Mach. Learn. Res.*, 16(1):2023–2049, 2015.
[72] Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.
[73] Jianqiang Wan, Sibo Song, Wenwen Yu, Yuliang Liu, Wenqing Cheng, Fei Huang, Xiang Bai, Cong Yao, and Zhibo Yang. Omniparser: A unified framework for text spotting key information

extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15641–15653, 2024.

[74] Yibo Wang, Guangda Huzhang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Lijun Zhang. SPACE: Noise contrastive estimation stabilizes self-play fine-tuning for large language models. In *Advances in Neural Information Processing Systems 39*, 2025.

[75] Yibo Wang, Hai-Long Sun, Guangda Huzhang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Lijun Zhang. Triplets better than pairs: Towards stable and effective self-play fine-tuning for LLMs. In *Advances in Neural Information Processing Systems 39*, 2025.

[76] Zhiruo Wang, Haoyu Dong, Ran Jia, Jia Li, Zhiyi Fu, Shi Han, and Dongmei Zhang. Tuta: Tree-based transformers for generally structured table pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1780–1790, 2021.

[77] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*, 2024.

[78] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[79] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[80] Bohao Yang, Yingji Zhang, Dong Liu, André Freitas, and Chenghua Lin. Does table source matter? benchmarking and improving multimodal scientific table understanding and reasoning. *arXiv preprint arXiv:2501.13042*, 2025.

[81] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[82] Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, Qi-Le Zhou, and De-Chuan Zhan. A closer look at deep learning on tabular data. *arXiv preprint arXiv:2407.00956*, 2024.

[83] Han-Jia Ye, Si-Yang Liu, and Wei-Lun Chao. A closer look at tabpfn v2: Strength, limitation, and extension. *arXiv preprint arXiv:2502.17361*, 2025.

[84] Han-Jia Ye, Huai-Hong Yin, De-Chuan Zhan, and Wei-Lun Chao. Revisiting nearest neighbor for tabular data: A deep tabular baseline two decades later. In *International Conference on Learning Representations*, 2025.

[85] Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. Large language models are versatile decomposers: Decomposing evidence and questions for table-based reasoning. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 174–184, 2023.

[86] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

[87] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023.

[88] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv e-prints*, pages arXiv–2407, 2024.

[89] Xuanliang Zhang, Dingzirui Wang, Longxu Dou, Qingfu Zhu, and Wanxiang Che. A survey of table reasoning with large language models. *Frontiers of Computer Science*, 19(9):199348, 2025.

[90] Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Binghong Wu, Lei Liao, Shu Wei, Yongjie Ye, Hao Liu, Wengang Zhou, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. In *Advances in Neural Information Processing Systems*, pages 7185–7212, 2024.

[91] Yilun Zhao, Boyu Mi, Zhenting Qi, Linyong Nan, Minghao Guo, Arman Cohan, and Dragomir Radev. Openrt: an open-source framework for reasoning over tabular data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 336–347, 2023.

[92] Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 9102–9124, 2024.

[93] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

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

[95] Fengbin Zhu, Ziyang Liu, Fuli Feng, Chao Wang, Moxin Li, and Tat Seng Chua. Tat-llm: A specialized language model for discrete reasoning over financial tabular and textual data. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 310–318, 2024.

[96] Wenwen Zhuang, Xin Huang, Xiantao Zhang, and Jin Zeng. Math-puma: Progressive upward multimodal alignment to enhance mathematical reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 26183–26191, 2025.

A Training Details

In this section, we will introduce some training details about our experiment, including the datasets we use and the hyperparameters in our experiment.

A.1 Dataset Details

Following HIPPO [40], we use Table Question Answering (TQA) and Table Fact Verification (TFV) tasks for training and evaluation.

All data statistics are shown in Table 2. we use five representative datasets: TABMWP [41], WTQ [52], HiTab [12], and TAT-QA [94]. We export FeTaQA [45] due to its evaluation metric being BLEU [51], which does not focus on the accuracy of question answering. For Table Fact Verification (TFV) tasks, we include TabFact [10] and InfoTabs [23]. To further assess the robustness of our approach, we conduct additional experiments on MMMU [86], evaluating our model on all table-related questions within this challenging multimodal benchmark.

Table 2: Dataset Statistics.

Dataset	Train	Test
Question Answering		
TABMWP	30,745	7,686
WTQ	17,689	4344
HiTab	8,941	1,586
TAT-QA	5,920	772
Fact Verification		
TabFact	31,321	6,845
InfoTabs	18,383	5,400
MMMU	-	165

The construction of the TURBO training dataset utilized TABMWP, WTQ, TAT-QA, TabFact, and InfoTabs. The HiTab dataset is excluded because it involves multi-level tables, which present formatting challenges when converted to Markdown. From each of the chosen datasets, we randomly extract 2,000 instances, resulting in a combined dataset of 10,000 training instances. Each instance consists of a structured table, a question, and the corresponding answer, which are then fed into a structure-aware reasoning tracer generator to generate high-quality reasoning traces. Then, reject sampling is used to further filter the data, resulting in 9,000 training instances in total. This dataset forms the foundation for supervised fine-tuning and subsequent reinforcement learning stages in the TURBO framework.

A.2 Hyperparameters.

As shown in Table 3, we provide the training hyperparameters for TURBO. Throughout all stages of training, we pre-train for one epoch in 9k generated data, with a batch size of 128.

During the SFT (supervised fine-tuning) stage, we adopt a learning rate of 2e-6 with a cosine decay schedule, a per-GPU batch size of 2, and 16 gradient accumulation steps across 4 A100-80G GPUs. The model is trained with bf16 precision and gradient checkpointing enabled. The maximum text and multimodal input lengths are set to 1500 and 4096 tokens, respectively.

In the RL stage, we fine-tune using a lower learning rate of 5e-7 with a constant schedule. The setup involves 8 A100-80G GPUs with a per-GPU batch size of 1 and the same accumulation steps. Each input prompt generates 16 candidate completions, and we apply a CLIP range of 0.2 to stabilize reward scaling. The sequence lengths are extended to 2000 (text) and 4596 (multimodal). We retain the AdamW optimizer and a warmup ratio of 0.1 across both stages, with no weight decay.

Table 3: Training hyperparameters.

Config	SFT Stage	RL Stage
Epoch	1	
Optimizer	AdamW	
Learning rate	2e-6	5e-7
Learning rate scheduler	Cosine	Constant
Weight decay	0.0	
Warmup ratio	0.1	
Text max length	1500	2000
Multimodal max length	4096	4596
Generation number	-	16
CLIP range	-	0.2
Batch size per GPU	2	1
Gradient accmulation steps	16	
GPU	4 × A100-80G	8 × A100-80G
Precision	Bf16	
Gradient checkpoint	True	

B More Experiments

In this section, we present additional experimental results to further validate the effectiveness of TURBO. We also include qualitative examples to illustrate the model's behavior and capabilities across various scenarios.

B.1 Comparison with Reasoning Models

We conducted additional experiments to compare our 8B TURBO model against leading large-scale, reasoning-focused models, even those that operate on structured text inputs, which is an inherently easier task. We compared TURBO with DeepSeek-R1 [21] (671B), a state-of-the-art reasoning LLM, and QvQ [54] (72B), another powerful model designed by Qwen team known for its reasoning capabilities. We also show some text-only results for reference based on our model. The result is shown in Table 4.

Table 4: Comparison with DeepSeek and the text-only results over datasets.

Method	Question Answering				Fact Verification		Average
	TABMWP	WTQ	HiTab	TAT-QA	TabFact	InfoTabs	
DeepSeek-R1	98.31	85.03	82.86	86.79	91.61	79.46	87.34
QvQ	96.38	79.54	79.03	78.26	93.25	74.91	83.56
Baseline-textonly	94.52	62.44	75.11	66.96	82.04	72.43	75.58
TURBO-textonly	96.29	66.50	71.31	73.21	87.74	82.44	79.58
TURBO	96.75	67.80	72.15	73.21	85.81	81.89	79.60

This comparison leads to several important conclusions:

Remarkable Performance Despite Modality Disadvantage: TURBO operates on raw table images, which requires it to first solve the complex perception problem of parsing the table structure and content before it can even begin to reason. In contrast, models like DeepSeek-R1 are given clean, structured Markdown tables—the very privileged information we use only during training. Despite this significant disadvantage in input modality, our 8B model achieves performance that is highly competitive with a 671B model, and even surpasses DeepSeek-R1 on the InfoTabs dataset. Our model also surpasses QvQ on TABMWP and InfoTabs datasets with smaller model size, which is a larger and stronger reasoning model.

Exceptional Parameter Efficiency: Our TURBO framework achieves these strong results with a model that is nearly 85 times smaller than DeepSeek-R1 (8B vs. 671B). This demonstrates that our proposed training methodology is highly effective at distilling complex reasoning abilities into a

much more efficient model. The strength of TURBO is not in model scale, but in its novel technique for bridging the modality gap and transferring reasoning skills.

Validation of our Core Claim: This comparison strongly validates our central thesis: by leveraging privileged information during training, we can equip a moderately-sized MLLM with reasoning capabilities on visual tables that approach, and in some cases exceed, those of massive LLMs working with clean, structured data.

Furthermore, we evaluate the generalization capability of our approach by comparing performance under a text-only setting. Specifically, we run both baseline and our TURBO model using structured tables as input. Results show that TURBO, though primarily trained in a multimodal setting, achieves strong performance even when only textual input is provided. The comparable results between TURBO and its text-only variant indicate that our training strategy successfully distills and transfers structured knowledge from tables into the model. This highlights the broader applicability of our approach and its ability to enhance both visual and textual reasoning in a unified framework.

B.2 Extension of TURBO

As a well-established model, Ovis2 provides a clean and controlled baseline, allowing us to isolate and clearly measure the specific performance gain contributed by our TURBO framework. By applying our methodology to a solid but not top-performing model, we can more transparently demonstrate the effectiveness of our privileged information training and two-stage learning process itself, rather than relying on the sheer power of an underlying SOTA model.

However, to directly address the question about the effectiveness of our framework and to test the generalizability of our approach, we conducted a new experiment applying the full TURBO training pipeline to the stronger Qwen2.5-VL model. The results in Table 5 are highly encouraging and strongly validate the robustness of our framework.

Table 5: Extension of our TURBO framework to Qwen2.5-VL.

	TABMWP	WTQ	HiTab	TAT-QA	TabFact	InfoTabs
Qwen2.5-VL	92.48	65.85	67.09	70.54	83.01	77.91
Qwen2.5-VL-SFT	96.02	69.77	69.73	72.59	85.92	78.17
Qwen2.5-VL-SFT-GRPO	96.92	70.35	71.64	71.55	87.27	79.93

This experiment leads to two important conclusions: 1) Applying our framework to the already stronger Qwen2.5-VL model yields a consistent and even slightly larger absolute performance gain. This demonstrates that our method of leveraging privileged information provides a fundamental enhancement to tabular reasoning capabilities, regardless of the underlying base model's initial strength. 2) By combining our TURBO framework with a more powerful base model, we achieve better results. This result showcases the scalability of our approach and its potential to push the boundaries of multimodal table reasoning even further when applied to future, more powerful base models.

B.3 Inference efficiency

Inference efficiency is a critical aspect of any deployable model. We have analyzed the inference speed of TURBO and can provide the following clarification.

Our two-stage training process (SFT + RL) only updates the weights of the model; it does not alter the model's architecture or increase its parameter count. Therefore, the fundamental per-token generation speed of our TURBO model is identical to that of the original Ovis2 baseline. The model itself is not inherently slower.

The observable increase in total inference time for TURBO comes from a deliberate design choice: the model generates a longer output sequence. Instead of producing only a final answer, TURBO generates a full reasoning trace. This naturally takes more time because more tokens need to be generated.

To provide concrete data, we measured the average inference time and output length on the WTQ dataset in Table 6:

Table 6: Inference time in WTQ dataset.

Model	Avg. Inference Time per Sample (A100-80G)
Ovis2 (Baseline)	0.8 seconds
TURBO	2.7 seconds
QvQ-72B	108 seconds

This data highlights a clear trade-off: TURBO's inference time is longer than its own baseline's because it's performing and articulating a complex reasoning process. This is the cost of achieving significantly higher accuracy and providing a traceable output for error analysis.

Competitive Speed Compared to Other SOTA Models: Crucially, when we compare TURBO's inference speed to that of other powerful, state-of-the-art multimodal models like QvQ, we find that our inference time is not an outlier and remains highly competitive. Despite generating a detailed reasoning trace, TURBO's speed is well within the acceptable limits for models in this performance class. This demonstrates that the efficiency cost for our significant accuracy gains is reasonable and practical.

In summary, the two-stage training does not slow down the model's intrinsic speed. The longer inference time is a direct and predictable consequence of generating more detailed, reasoned outputs. This represents a favorable trade-off for achieving state-of-the-art accuracy, and the resulting speed is comparable to other leading models.

B.4 More Cases

We present more detailed cases in Figure 6. These cases illustrate the diverse reasoning types and challenges, including multi-step arithmetic, compositional conditions, and fine-grained table comprehension. These cases demonstrate that our TURBO can do well in both visual grounding, mathematics, and logical reasoning.

C Limitations and Future Works

Failure cases. Based on a thorough review of our model's outputs, we have identified two primary categories of failure cases.

OCR Errors in Content Extraction: The first step in visual table reasoning is accurately extracting the textual content. Like most MLLMs, TURBO is susceptible to occasional OCR errors, especially with dense text, small fonts, or image artifacts. These are often subtle but critical. For example, we observed cases where a number like "30001" was misidentified as "3001" or a keyword was misread. Such low-level perception errors inevitably lead to incorrect results in subsequent reasoning steps, as the model is operating on faulty input data.

Errors in Complex Numerical or Logical Reasoning: Even when all table data is extracted perfectly, TURBO can sometimes fail at the final reasoning stage. This is a known challenge for large language models in general. We observed that these errors are most common in tasks requiring multi-step arithmetic calculations or complex logical deductions. For instance, the model might correctly formulate a plan to sum several cells and then subtract a value, but make a mistake during the final calculation. This indicates that while our method significantly enhances reasoning capabilities, it does not completely eliminate the inherent fallibility of LLMs in complex computation.

Limitations. While this work focuses on scenarios where structured tables are unavailable at test time—requiring inference purely from table images—there exist many practical situations where both the image and its underlying structured representation can be accessed. Exploring how to effectively incorporate structured tables as additional inputs during inference presents a promising direction for future research.

Another limitation lies in the type of tables considered. Most of our experiments involve relatively clean and regular tables, where each row corresponds to a distinct entity and each column to a specific

Question: A candy dispenser put various numbers of orange candies into bags.
How many bags had at least 22 orange candies but fewer than 86 orange candies?

Orange candies per bag

Stem	Leaf
1	6 8 9
2	1 6 9
3	9
4	1
5	2 2 5 6 8
6	1 2
7	
8	9

Turbo:<think>Okay, let's see. The problem is asking how many bags had **at least 22** orange candies **but fewer than 86**. The data is given in a stem-and-leaf plot.
First, I need to understand how stem-and-leaf plots work. The stem represents the tens place, and the leaf is the ones place. So, for example, a stem of 1 with leaves 6, 8, 9 would be 16, 18, 19.
Looking at the table, … the stems that fall within this range are 2, 3, 4, 5, 6, and 7. Stem 8 is 89, which is above 86, so that's excluded.
Now, let's count the leaves for each of these stems:\n\n- Stem 2: Leaves are 1, 6, 9. That's 21, 26, 29, but 21 should be excluded. So 2 bags …
Adding these up: **2 + 1 + 1 + 5 + 2 = 11**. So the total number of bags with at least 22 but fewer than 86 candies is 11.</think>
<answer>11</answer>

Sample No.	*Moisture Content*	*Bulk Density (g/cc)*
1	4.8	2.164
2	7.5	2.322
3	7.8	2.351
4	8.9	2.345
5	9.7	2.327

Question: The results of a compaction test on samples of soil that are to be used for an embankment on a highway project are listed below. Determine the optimum moisture content.
A: 10%.
B: 8%.
C: 9%.

Turbo: <think>Okay, let's see. The question is asking for the optimum moisture content based on the compaction test results provided in the table. The table lists five samples with their respective moisture content and bulk density.
First, I need to understand what \"**optimum moisture content**\" means. In soil compaction, the optimum moisture content is the percentage at which the soil can be compacted to achieve the maximum dry density. …
Looking at the table, the samples are numbered 1 to 5. The moisture content increases from **4.8% to 9.7%.** The bulk density values also increase, **starting from 2.164 g/cc and going up to 2.351 g/cc**. …
The key here is to find the sample where the bulk density is highest, as that would correspond to the optimum moisture content. Let's check the bulk density values:\n\n- Sample 1: ….\n\nSample 3 has the highest bulk density at 2.351 g/cc, which is the maximum value in the table. …
Therefore, the optimum moisture content is 7.8%, which is closest to option B: 8%.</think>
<answer>B</answer>

Figure 6: More cases of Turbo.

attribute. However, real-world tables are often more diverse and complex. For example, datasets like HiTab and MMMU include tables with merged cells, nested headers, and irregular layouts that challenge current MLLMs. A more systematic analysis of how different table structures affect multimodal reasoning performance would be a valuable extension of this work.

Future work. Lastly, while our proposed TURBO framework significantly improves the reasoning abilities of open-source MLLMs, a substantial performance gap remains when compared with state-of-the-art proprietary LLMs such as DeepSeek. We hope this paper inspires the community to further explore real-world multimodal reasoning scenarios and to continue narrowing the gap between open- and closed-source models, contributing to more robust and accessible MLLMs.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction include the claims made in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in Section C.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: There are no theoretical results in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All information needed to reproduce the main experimental results of the paper is available and easy to follow.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The data is available. The code is easy to follow according to existing repositories.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All training details are in the Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The evaluation of LLM and MLLMs does not report error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information is provided in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Answer: [NA]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning, especially MLLMs and tabular reasoning. There are many potential societal consequences in our work, none of which we feel must be specifically highlighted here.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The models and datasets here are all public and frequently used.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the original papers that produced the code package and datasets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve potential risks incurred by study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: This paper provide a new method for multimodal large language models in tabular reasoning. Details can be found in Section 4 and Section 5.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.