TABLET: A LARGE-SCALE DATASET FOR ROBUST VISUAL TABLE UNDERSTANDING

Anonymous authors

Paper under double-blind review

ABSTRACT

While table understanding increasingly relies on pixel-only settings where tables are processed as visual representations, current benchmarks predominantly use synthetic renderings that lack the complexity and visual diversity of real-world tables. Additionally, existing visual table understanding (VTU) datasets offer fixed examples with single visualizations and pre-defined instructions, providing no access to underlying serialized data for reformulation. We introduce TABLET, a large-scale VTU dataset with 4 million examples across 20 tasks, grounded in 2 million unique tables where 88% preserve original visualizations. Each example includes paired image-HTML representations, comprehensive metadata, and provenance information linking back to the source datasets. Fine-tuning vision-language models like Qwen2.5-VL-7B on TABLET improves performance on seen and unseen VTU tasks while increasing robustness on real-world table visualizations. By preserving original visualizations and maintaining example traceability in a unified large-scale collection, TABLET establishes a foundation for robust training and extensible evaluation of future VTU models.¹

1 Introduction

The field of table understanding focuses on techniques for representing and interpreting tabular data to support a wide range of practical tasks such as question answering, summarization, and information extraction. Research in this area has traditionally representated tables as structured text, encoding their content and layout through linearized or graph-based representations (see Figure 1b; Herzig et al. 2020; Zhang et al. 2020; Liu et al. 2022). While this unimodal view remains effective in certain domains, many tables found in documents and webpages contain irregular structures, rely on visual formatting (e.g., merged cells, background colors, font variations), or embed multimodal elements such as images (see Figure 1a). Advances in Vision-Language Models (VLMs; Radford et al. 2021; Liu et al. 2023) have provided impetus for treating tables as images, eschewing the step of rendering them as text sequences (like Markdown or HTML). The conceptual simplicity of this approach, coupled with improved performance on several tabular tasks (Alonso et al., 2024; Zhou et al., 2025) has driven significant research interest (Zheng et al., 2024b; Su et al., 2024; Jiang et al., 2025) in Visual Table Understanding (also known as Multimodal Table Understanding). Visual representations of tables are not only merely convenient but in many cases necessary, particularly for VLM agents that interact with the world exclusively through pixels (e.g., on a screen) and must interpret tables directly in their visual form (Deng et al., 2023; Zheng et al., 2024a; Lu et al., 2024).

Despite the growing relevance of VTU, there are few resources that support training models *directly* on image-based representations of tables. Existing benchmarks like MMTab (Zheng et al., 2024b) consist of web tables (e.g., from Wikipedia which is a common source for many tabular datasets), that are serialized and subsequently rendered as synthetic images (see Figures 1b,c). These images do not preserve the original visual characteristics of the tables and do not reflect the diversity and complexity of real-world layouts. As a result, models trained on such data face a train-test mismatch, since the visual patterns learned from serialized renderings do not generalize well to naturally occurring tables failing to capture critical visual cues like subtle ruling lines, intricate merged cell layouts, background colors, font variations, or embedded images that are inherent to real-world table comprehension (compare Figure 1a and 1c). An exception is WikiDT Shi et al. (2024), which provides

¹Dataset, code, and other resources are available at xxx.yyy.zzz.

056

058

060

061

062

063

064

065

066

067 068 069

071

073

074

075

076

077

079

081

083

084

085

087

090

092

093

095

097 098

099

100

101

102

103

104

105

106

107

Figure 1: Previous datasets render table images from serialized tables, losing original visual details. In contrast, TABLET locates and retrieves the original table visualizations across 14 tabular datasets, resulting in 4M examples grounded in 2M unique tables.

access to original table visualizations from web pages but focuses exclusively on table question answering. Likewise, datasets designed for processing PDF documents or screenshots (Zhong et al., 2020; Lu et al., 2023) preserve original table visualizations but support only a single task.

In this work, we introduce TABLET², a large-scale dataset designed to enable vision-language models to learn generalizable skills for table understanding by leveraging visual features from table images. Unlike existing datasets that focus on a single task or use synthetic renderings, TABLET preserves lossless representations of real-world table images and spans a diverse set of tasks, including tableto-text generation (Parikh et al., 2020), table fact verification (Chen et al., 2020b), and table-based question answering (Pasupat & Liang, 2015). TABLET facilitates training and evaluation on table formats that align with those encountered in downstream applications, particularly in pixel-based agents that process visual input directly. It contains 4 million examples across 20 tasks, derived from 2 million unique tables, 88% of which retain their original visualization. To promote flexibility and foster future research, we provide both image and HTML representations of tables, together with metadata and links to the source datasets. Pairing table images with HTML enables models to learn from naturally occurring and synthetic table images when the former are not available. TABLET's large-scale and task aggregation provide a unique, foundational resource that complements research integrating textual and visual modalities for training and reasoning (Liu et al., 2025; Jiang et al., 2025). We experimentally demonstrate that fine-tuning on TABLET improves performance on inand out-of-domain benchmarks. Our work makes the following contributions:

- TABLET addresses a critical gap in large-scale, visually faithful datasets for VTU, with 4 million examples and 2 million tables, most of which are preserved in their original form.
- We show that supervised fine-tuning on TABLET enhances VLM performance across seen
 and unseen table understanding tasks. We release a VLM model trained on TABLET that
 achieves state-of-the-art results on 10 out of 14 table understanding benchmarks.
- TABLET is designed for extensibility: all tasks are provided in a unified format, paired with HTML table serializations that can be rendered back into the original visualization, along with example metadata and links to the source datasets to facilitate reuse and reformulation.

2 Related Work

Early work approached the problem of table understanding from a unimodal perspective, treating tables as structured text and developing general-purpose models that generalized across multiple tasks (Li et al., 2023; Zhang et al., 2024). Recent progress in natural image understanding afforded by increasingly better VLMs (Radford et al., 2021; Liu et al., 2023) has led to development of models that can successfully process visually rendered text beyond natural imagery and perform table understanding holistically, within a multimodal framework (Kim et al., 2022; Lee et al., 2023; Zhang et al., 2023; Ye et al., 2023; Hu et al., 2024; Alonso et al., 2024; Jiang et al., 2025; Su et al., 2024; Zhao et al., 2024; Zheng et al., 2024b; Su et al., 2024).

²Our dataset is called TABLET partly as a nod to the Scottish sugary confection which is often cut into uniform squares or rectangles, resembling a table.

To this effect, several benchmarks have been proposed recently to evaluate progress in VTU. For instance, TableVQA-Bench (Kim et al., 2024) focuses on visual table question answering but relies on synthetic table images. In contrast, MMTBench (Titiya et al., 2025) includes original table visualizations, along with visually rich images and interleaved charts. While both are valuable for evaluation, they are limited in scope and do not provide large-scale training data, restricting their utility for developing generalizable models.

Larger training datasets (with training splits) have historically focused on image-based Table Structure Recognition (TSR), including PubTabNet (Zhong et al., 2020), TableBank (Li et al., 2020), and TabComp (Gautam et al., 2025). Beyond TSR, Alonso et al. (2024) created image-rendered versions of existing datasets like ToTTo (Parikh et al., 2020), relying on lossy renderings that discard visual features. WikiDT (Shi et al., 2024) preserves original table visualizations but only targets Visual TableQA. MMTab (Zheng et al., 2024b) is a large VTU dataset, with 150k TSR pre-training examples and 232k instruction examples across 19 tasks. While MMTab is a useful resource that we build upon, it relies on synthetic renderings of serialized tables, lacks traceability to original data sources, and remains limited in size compared to the scale needed for general-purpose VTU.

In this work, we do not introduce a task-specific dataset or model. Instead, we create a large-scale resource aimed at enhancing table understanding in general-purpose vision-language models like Gemma-3 (Team et al., 2025) or Qwen2-VL (Wang et al., 2024), under the assumption that tables for many tasks are seen as images, and therefore can be naturally processed by these models. This distinguishes TABLET from many recent efforts that introduce task-specific datasets or specialized architectures as it is designed to be a holist resource for advancing VLM capabilities for tables.

3 THE TABLET DATASET

3.1 Overview

TABLET is a large-scale dataset that aggregates a total of 4,066,545 examples, combining 20 TU tasks sourced from 14 datasets (see Section 3.3 for a breakdown of examples per task). TABLET includes 2,031,256 unique table images with visualizations from Wikipedia (61.5%), Pub-TabNet (25.1%; Zhong et al. 2020), TabMWP (1.9%; Lu et al. 2023, and synthetically rendered images (11.5%) from serialized tables (see Section 3.2 for details).

Examples in TABLET are framed as instructions, leveraging the benefits of instruction-tuning for large language models. TABLET does not introduce new data instances; rather it repurposes examples from existing datasets, referred to as *seed* datasets, rephrasing their tasks into an instruction format. Examples are drawn from 14 seed datasets: TURL (Deng et al., 2020), ToTTo (Parikh et al., 2020), TabFact (Chen et al., 2020a), WikiTableQuestions (Pasupat & Liang, 2015), HybridQA (Chen et al., 2020b), HiTab (Cheng et al., 2022), PubTabNet (Zhong et al., 2020), TabMWP (Lu et al., 2023), TAT-QA (Zhu et al., 2021), InfoTabs (Gupta et al., 2020), WikiBIO (Lebret et al., 2016), FeTaQA (Nan et al., 2022), MMTab Zheng et al. (2024b), and DocStruct4M (Hu et al., 2024).

Datasets that include only images and fixed prompts restrict researchers to the provided visualizations and instruction formulations, limiting the exploration of new prompting techniques or the use of table data to create additional examples. To avoid this, each example in TABLET is provided in a unified format, including the instruction used in this work, its corresponding atomic data, the HTML version of the table, the original source ID, and additional metadata (detailed in Appendix7.4).

3.2 Image sources

Wikipedia Tables Wikipedia tables generally follow a hierarchical table structure and are often rich in visual information. 61.5% of the tables in TABLET are lossless visualizations sourced from Wikipedia (See Appendix 7.2 for a breakdown of image sources per task). These tables are referenced in 67.4% of all examples in the dataset. Each table from the seed datasets was traced back to its original visualization in the corresponding Wikipedia article and captured in a screenshot.³. A key challenge in retrieving the original visualizations was that the seed datasets were created at

³Table images were rendered using Firefox in headless mode (version 142.0.1 with GeckoDriver version 0.36.0) at an effective rendering density of 96 PPI (which stands for Pixels Per Inch).

Task	Seeds	Train	Dev	Test
ent_link	TURL	1,236,128 (35.3%)	74,282 (35.4%)	213,494 (60.8%)
col_type	TURL	602,406 (17.2%)	13,188 (6.3%)	12,802 (3.6%)
struct_aware_parse	M^3	513,482 (14.6%)	9,115 (4.3%)	1,102 (0.3%)
wikibio	WikiBIO	582,659 (16.6%)	72,831 (34.7%)	72,831 (20.7%)
hybridqa	HybridQA	62,670 (1.8%)	3,466 (1.7%)	3,463 (1.0%)
fetaqa	ToTTo	3,006 (0.1%)	577 (0.3%)	1,079 (0.3%)
hitab	NSF, StatCan, ToTTo	7,417 (0.2%)	1,670 (0.8%)	1,584 (0.5%)
infotabs	InfoTabs	16,538 (0.5%)	1,800 (0.9%)	5,400 (1.5%)
tabfact	TabFact	87,717 (2.5%)	12,389 (5.9%)	12,326 (3.5%)
tabmwp	TabMWP	23,059 (0.7%)	7,686 (3.7%)	7,686 (2.2%)
tat-qa	TAT-QA	2,201 (0.1%)	278 (0.1%)	277 (0.1%)
totto	ToTTo	110,934 (3.2%)	7,077 (3.4%)	7,084 (2.0%)
wikitq	WikiTableQuestions	14,152 (0.4%)	3,537 (1.7%)	4,344 (1.2%)
rel_extraction	TURL	60,615 (1.7%)	2,145 (1.0%)	2,030 (0.6%)
table_instruction	\mathbf{M}^1	136,944 (3.9%)	0 (0.0%)	0 (0.0%)
row_column_extraction	\mathbf{M}^1	7,721 (0.2%)	0 (0.0%)	957 (0.3%)
table_cell_extraction	\mathbf{M}^1	7,727 (0.2%)	0 (0.0%)	966 (0.3%)
table_cell_location	M^1	7,708 (0.2%)	0 (0.0%)	956 (0.3%)
table_recognition	M^1	6,927 (0.2%)	0 (0.0%)	912 (0.3%)
table_size_detection	\mathbf{M}^1	7,800 (0.2%)	0 (0.0%)	950 (0.3%)
merged_cell_detection	\mathbf{M}^2	7,500 (0.2%)	0 (0.0%)	950 (0.3%)
Total		3,505,311	210,041	351,193

Table 1: TABLET splits, tasks, and seed datasets; M¹: InfoTabs, NSF, StatCan, TabMWP, TAT-QA, TabFact, ToTTo, WikiBIO, WikiTableQuestions; M²: InfoTabs, NSF, StatCan, TAT-QA, TabFact, ToTTo, WikiBIO; M³: PubTabNet, TabFact, WikiTableQuestions. Proportion of examples in each set (train/dev/test) is shown in gray within parentheses.

different times, while Wikipedia articles are continuously updated. To address this, we relied on the metadata released with the seed datasets and also reached out to their authors to find out when the tables were harvested (see Appendix 7.3). We used Wikipedia's archiving API to retrieve each article as it appeared at the time of crawling. All Wikipedia tables in TABLET are linked to their source via both the page identifier and the corresponding revision identifier (*oldid*). As Wikipedia pages often contain multiple tables, we needed to identify which one corresponded to the seed example. We did this by computing the Levenshtein (1966) edit distance between Wikipedia tables and seed tables, both represented in markdown format, and selecting those with the highest similarity value. We set a minimum similarity threshold of 0.70; in cases without a match, the serialized seed table was rendered as an image.

Synthetic Tables Wikipedia tables that could not be retrieved, either due to low similarity scores or other issues, were synthetically generated by converting their serialized format to HTML and rendering them using the same browser configuration used for the original Wikipedia tables. These synthetic tables account for 11.5% of the dataset, contributing 747,002 additional examples (18.4% of the total). In Section 5, we show that combining original table images with synthetic visualizations during training allows for an increase in example count and visualization diversity that leads to better model performance compared to using synthetic or original tables alone.

Document Tables We also incorporate tables from other domains to improve model generalization. These include tables from scientific articles in PubTabNet (Zhong et al., 2020) and rendered tables from TabMWP (Lu et al., 2023). PubTabNet contributes 509,892 tables to our dataset (25.1% of the total), accounting for 97.5% of the structure-aware learning task incorporated from DocStruct4M (Hu et al., 2024). TabMWP adds 38,431 tables (1.9% of the total); while smaller in scale, TabMWP offers valuable coverage of numerical reasoning examples.

3.3 TABLE UNDERSTANDING TASKS

TABLET aggregates a total of 20 TU tasks grouped into 7 broad categories: Table Interpretation, Question Answering, Table-to-Text Generation, Table Numerical Reasoning, Table Natural Lan-

guage Inference (NLI), and Table Structure Understanding. Subtasks within each category require different skills and demonstrate various degrees of complexity, ranging from basic table structure understanding to downstream tasks that combine tabular reasoning with other skills, such as information retrieval, numerical reasoning, or natural language inference. Table 1 provides a breakdown of examples per task and their source dataset. We briefly describe each task below and provide more details in Appendix 7.5.

Table Interpretation comprises three tasks, namely *Column Type Annotation*, *Entity Linking*, and *Relation Extraction* (ent_link, col_type, and rel_extraction in Table 1). Sourced from TURL (Deng et al., 2020), these tasks do not represent downstream applications but target basic table understanding skills essential for tackling more complex tasks. All instructions for these tasks were aggregated from TURL. They constitute the largest group in our training set, with 1,899,149 examples (54.2%).

Table Question Answering is represented by *Free-form Table QA* (FeTaQA, Nan et al. 2022), *Hierarchical Table QA* (HiTabQA; Cheng et al. 2022), *Hybrid QA* (HybridQA; Chen et al. 2020b), and *Table QA* (WikiTableQuestions; Pasupat & Liang 2015). These tasks range from simple question answering over table content to multi-hop reasoning that combines textual and tabular information. They also vary in terms of table complexity and the expected answer format. Question answering represents 2.5% of TABLET's training set (87,245 examples).

Table-to-Text Generation is exemplified by two subtasks, namely *Highlighted Cell Text Generation* (ToTTo; Parikh et al. 2020) and *Wikipedia Biography Generation* (WikiBio; Lebret et al. 2016). Both tasks involve generating text based on table content, concise biographies in the case of WikiBio, and short descriptions based on visually highlighted cells, in the case of ToTTo. Together, they account for 693,593 training examples (19.8%).

Table Numerical Reasoning combines *Tabular Math Word Problem Solving* (TabMWP; Lu et al. 2023) and *Hybrid-context Financial QA* (TAT-QA; Zhu et al. 2021). Given a table and a mathematical question, the model must perform reasoning over table values. This type of numerical reasoning represents 0.8% of TABLET's training set (25,260 examples).

Table NLI includes two entailment tasks: *Table Fact Verification* (TabFact; Chen et al. 2020a) and *Infobox Natural Language Inference* (InfoTabs; Gupta et al. 2020) where statements as supported or refuted based on table content. They represent 104,255 examples (3%) in TABLET's training set.

Table Structure Understanding is an umbrella category for tasks designed to facilitate structural understanding of tables: *Merged Cell Detection, Row and Column Extraction, Table Cell Extraction, Table Cell Location, Table Recognition, Table Size Detection*, and *Structure-aware Parsing* in Table 1). Examples for the first seven tasks are aggregated from MMTab, while Structure-aware Parsing is derived from DocStruct4M (Hu et al., 2024). These tasks are based on tables from multiple seed datasets (see rows with seeds M², M¹, and M³), including InfoTabs (Gupta et al., 2020), NSF (National Center for Science and Engineering Statistics, 2019), StatCan (Statistics Canada, 2024), TabMWP (Lu et al., 2023), TAT-QA (Zhu et al., 2021), TabFact (Chen et al., 2020a), ToTTo (Parikh et al., 2020), WikiBio (Lebret et al., 2016), WikiTableQuestions (Pasupat & Liang, 2015), and PubTabNet Zhong et al. (2020). While we maintain the original MMTab instructions, we use our own table visualizations. This group contributes 558,865 examples (15.8%) to our training.

Instruction Following To facilitate evaluation, we fine-tune and evaluate all models using instructions that explicitly require the final answer to be encapsulated in a JSON object, regardless of any additional tokens the model may generate. To increase instruction diversity and mitigate catastrophic forgetting, we include a dedicated instruction-following set from MMTab. These instructions rephrase a subset of examples from the above tasks using a different template and do not require the JSON output format. While not a task in itself, this set adds instruction diversity and reduces overfitting to a single output style. We aggregate these examples directly from MMTab while using our visualizations. This set contributes 136,944 training examples (3.9%).

3.4 Linking to Sources and Highlighting

TABLET is designed with extensibility in mind. Existing datasets such as TableInstruct (Zhang et al., 2024), DocStruct4M (Hu et al., 2024), and MMTab (Zheng et al., 2024b) do not provide a clear reference to the original examples from which their instructions were derived. However, the absence of such pointers, makes reuse, modification, and augmentation difficult. Additionally,

image-based datasets such as MMTab lack serialized table versions in text format, providing only rendered images as representations. To address these issues, each example in TABLET includes both the identifier of the original example and the identifier of the corresponding table from the source dataset. Identifiers follow a simple, human-readable format, and the released code provides a function to retrieve the original examples given their IDs.

To support future extensions and new tasks, TABLET also provides serialized versions of its tables in HTML. For the tables successfully retrieved from Wikipedia, this HTML corresponds to the raw table object in the original HTML of the article. These tables can be rendered with their full visualization using Wikipedia's official CSS stylesheets or through the rendering functions provided with our code. The remaining tables are either provided as HTML in the seed datasets or converted into HTML from their serialized representations.

This feature is particularly useful in tasks involving cell pointing or highlighting. Prior work has shown that models can achieve competitive results by relying uniquely on highlighted values (An et al., 2022; Alonso et al., 2024). We also observed in our experiments that as long as highlighted cell values are mentioned in the prompt, models are able to perform the task (e.g., Column Type Annotation, Entity Linking, Relation Extraction, FeTaQA, and ToTTo) regardless of the table provided. In other words, the model learns to ignore the table during training, which can be detrimental for tasks requiring genuine table understanding. We mitigate this, by exploiting the HTML serialization to directly locate the highlighted cells in the source tables. By cross-referencing the highlighted spans from the seed datasets with the actual table content, we generate versions of the tables with explicit highlights while preserving the original visual features.

3.5 COMPARISON WITH MMTAB

MMTab (Zheng et al., 2024b) is also a dataset for visual table understanding; however, unlike TABLET, it does not include the full set of training examples from the original datasets (e.g., only 12.4% of ToTTo compared to our 91.8%; see Appendix 7.6), and it provides no development splits. MMTab covers 20 tasks while TABLET covers 21 (inculding instruction tuning); four tasks are unique to the former and five to the latter, and overall our dataset is 9.36 times larger (comprising 4,066,545 examples compared to MMTab's approximate 433,376 examples). Moreover, only a few tasks in MMTab include identifiers that can be linked back to their source examples, limiting developers to the provided instruction text. While MMTab relies on synthetically rendered tables with predefined styles, which omit original formatting and embedded images, TABLET retains the original visualizations, enabling models to exploit a fuller range of visual features.

4 EXPERIMENTAL SETTING

Our experiments are motivated by two questions: (1) Does training VTU models with lossless table visualizations improve performance on VTU tasks? and (2) Is TABLET diverse enough so that supervised fine-tuning (SFT) improves generalization on unseen VTU tasks? We also examine whether training only with original visualizations performs better than mixing original and synthetic ones, whether a balanced task distribution is preferable to the full dataset distribution, and whether including various table interpretation tasks contributes to VTU performance.

Backbone Model All experiments employ the same backbone model, namely Qwen2.5-VL with 7B parameters (Qwen et al., 2025). This model provides a good tradeoff between performance and computational requirements. Together with InternVL3 (Zhu et al., 2025), it was the best-performing open-weight VLM in its class, setting a new benchmark for Single Page Document VQA (Mathew et al., 2021) at the time of writing. Exploring the best backbone model for VTU or advanced SFT strategies is outside the scope of this work. Full SFT configurations are detailed in Appendix 8; results for other open-weight models on TABLET (test set) are reported in Appendix 8.4.

Evaluation We evaluate fine-tuned models on eight *held-in* tasks: three focus on Table QA (WikiTableQuestions, HiTabQA, FeTaQA), two on text generation (WikiBio, ToTTo), two on Table Numerical Reasoning tasks (TabMWP, TAT-QA), and one on Table NLI (TabFact). To test generalization, we further evaluate on six *held-out* tasks: HybridQA, InfoTabs, TabMCQ (Jauhar et al., 2016), AIT-QA (Katsis et al., 2022), Table Recognition, and PubHealthTab (Akhtar et al., 2022). The last four are directly extracted from MMTab and use instructions generated with their own templates,

	WikiBio ToTTo		Wik	iTQ	TabMWP	HiTAB		TabFact		FeTaQA		TAT-QA		
Model	synth	org	synth	org	synth	org	org	synth	org	synth	org	synth	org	org
0-Shot	2.0	2.5	10.8	8.9	57.7	50.8	51.9	11.1	2.5	70.4	59.4	9.9	9.8	48.7
TABLET-B _{synth}	11.3	4.8	12.4	14.4	58.4	56.5	84.3	33.5	16.0	63.1	65.5	28.6	27.9	33.4
TABLET-B _{org}	9.6	11.5	13.7	15.9	58.5	56.2	84.2	35.1	24.7	62.1	63.7	28.4	28.4	33.0
$TABLET-B_{mix}$	10.3	12.0	12.5	14.1	58.9	56.5	84.4	38.5	26.9	68.2	70.0	28.6	28.5	52.5

Table 2: Comparison of **Qwen2.5-VL-7B-Instruct** models fine-tuned on different TABLET-B variants and in zero-shot mode; models are evaluated on original (lossless) table visualizations and synthetic ones across **held-in benchmarks**. Results for TabMWP and TAT-QA exclude the mixed setting, as all tables already use original visualizations. We report exact match accuracy for WikiTQ, TabMWP, TabFact, and TAT-QA; BLEU for WikiBio, ToTTo, and FeTaQA; and F1 for HiTab. ToTTo results are reported on its development set. Best model per setting (synth/org) shown in bold.

adding diversity in both table and instruction styles. Although included in TABLET, the 86,135 training examples from HybridQA, InfoTabs, and Table Recognition were excluded from fine-tuning to enable held-out evaluation. Models trained on the full dataset would potentially achieve even stronger results than those reported here.

5 RESULTS AND ANALYSIS

Lossless vs Synthetic Table Visualizations We first assess whether models benefit from being trained with original table visualizations, in particular when exposed to real-world tabular images rather than synthetic ones at test time. This experiment was conducted on a subset of TABLET corresponding to MMTab as this benchmark represents the current state of the art in multimodal table understanding. We refer to this subset as TABLET-BASE (TABLET-B). We replaced all Wikipediabased table images in MMTab with their original visualizations and kept tables that did not undergo lossy serialization unchanged (TabMWP, PubHealthTab, TabMCQ, AIT-QA); we also left MMTab's instructions unaltered. We refer to this partition of the dataset as TABLET-Borg and contrast it with TABLET-B_{synth} which is the same version but with synthetic table images. TABLET-B contains 238,980 examples (in both original and synthetic versions). We also create TABLET-B_{mix} which combines original with synthetic tables. Specifically, we extend TABLET-Borg with all examples whose original visualizations could not be retrieved, using our synthetic renders. This set contains 371,292 examples (43.5% original visualizations and 35.6% synthetic). We perform full fine-tuning (without LoRA) on Qwen2.5-VL-7B for 3 epochs and evaluate in the 8 held-in tasks and 4 held-out tasks mentioned above. These tasks contain either original visualizations or synthetic ones to simulate various test cases (e.g., a model has seen only synthetic/original table images but is tested on original/synthetic ones).

Our results in Table 2 (held-in tasks) show that lossless models outperform synthetic ones in two tasks (WikiBio, HiTab), match performance in four (WikiTQ, TabMWP, FeTaQA, TAT-QA), and underperform in one (TabFact). For reference, we also report the 0-shot performance of Qwen2.5-VL-7B. The fact that there is no pronounced difference between lossless and synthetic visualizations is perhaps not surprising since most benchmarks were originally designed around serialized, text-only representations, without paying heed to visual cues (TURL, WikiBIO, ToTTo, TAT-QA). Thus, while lossless visualizations contain more information, current benchmarks do not test whether models benefit from them. Developing benchmarks with high visual variability and challenging layouts would better highlight scenarios where lossless visualizations offer a distinct advantage.

When comparing models trained with lossless table visualizations (TABLET- B_{org}) against a mixture of lossless and synthetic (TABLET- B_{mix}), we find that mixed training improves results on four tasks (WikiBio, HiTab, TabFact, TAT-QA) and obtains equivalent performance on three (WikiTQ, TabMWP, FeTaQA). However, note that this improvement could be due to TABLET- B_{mix} having more samples compared to the other partitions. In general, we observe that model performance degrades when evaluating on original visualizations. This setting appears to be consistently more challenging, but robustness improves when adding original tables into the training (see TABLET- B_{mix} row in table 2). We measure degradation as the average percentage-point change in perfor-

	Held-in Datasets												
	WikiBio	ToTTo	WikiTQ	TabMWP	Hitab	TabFact	FeTaQA	TAT-QA					
0-Shot	2.5	9.1*	53.4*	59.1*	31.2*	73.9*	7.0*	6.9*					
1-Shot	4.4	16.1*	49.8*	57.7*	37.3*	72.8*	8.5*	10.1*					
MMTab	6.4*	12.6*	48.7*	80.4*	41.5*	57.0*	1.7*	9.4*					
TABLET-S	2.9	28.9	56.8	84.0	64.8	78.9	28.7	27.8					
TABLET-M	3.1	29.3	56.6	84.0	67.0	79.5	30.7	31.0					
TABLET-L	<u>3.8</u>	<u>30.4</u>	55.5	<u>84.5</u>	67.5	79.5	<u>31.5</u>	32.5					

Table 3: Evaluation on eight held-in datasets with **Qwen2.5-VL-7B-Instruct** in 0/1-shot mode and fine-tuned on MMTab and different TABLET sizes. Results for ToTTo correspond to evaluation on the development set. We report exact match accuracy for WikiTQ, TabMWP, TabFact, and TATQA; BLEU for WikiBio, ToTTo, and FeTaQA; and F1 for HiTab. Best performing model per task is shown in bold; we mark with * models significantly different (p < 0.05, using bootstrap resampling) from those trained on TABLET (highlighted in gray). Models that perform significantly better within the TABLET group are underlined.

							12.5	12.5	TabMWp InfoTabs	
	Infotobo		eld-out-Dat	tasets PubHealthTab	. I I volumi d O A	TakDaa	12.5	5.2	WikiTQ Other ColType	
0-Shot	64.5*	84.4	51.7*	64.7*	34.2*	24.5*	5.6	5.4	EnTLink Wikibio	
1-Shot	65.5*	61.0*	43.1*	70.0*	30.5*	27.9*	12.5	12.3	HybridQA	1
MMTab	56.9*	89.1	56.6*	63.9*	27.1	43.6*	9.9	11.3	TabFact Totto	
TABLET-S	57.2	88.2	58.9	64.7	22.8	43.9			Struct Aware Parse	
TABLET-M	61.5	88.3	62.4	66.3	27.7	44.1			TableInstruct	
TABLET-L	61.4	87.9	<u>70.8</u>	<u>70.2</u>	25.1	45.4	a.TABL	ЕТ-М		

RelExtraction

Table 4: **Left:** Evaluation on five held-out datasets with **Qwen2.5-VL-7B-Instruct** (0/1-shot) and fine-tuned on MMTab and different TABLET sizes. We report Tree-Edit- Distance-based Similarity for Table Recognition (TabRec) and accuracy for all other benchmarks. HybridQA results are reported on the development set. Best performing model per task is shown in bold; we mark with \ast models significantly different (p < 0.05, using bootstrap resampling) from those trained on TABLET (highlighted in gray). Models that perform significantly better within the TABLET group are underlined. **Right:** Distribution of examples per task in TABLET-L (b) and TABLET-M (a).

mance between original and mixed visualizations. The baseline model degrades the most, with a drop of 28.9 percentage points, particularly in tasks like HiTab and TabFact. The model trained on TABLET-B_{synth} degrades by 22.35 points, compared to 6.63 and 7.87 points for models trained on TABLET-B_{org} and TABLET-B_{mix}, respectively. The model trained purely with lossless tables is most robust, maintaining stable performance when exposed to both original and synthetic styles (a detailed breakdown of degradation per task and model is in Appendix 8.2). On held-out tasks models trained on TABLET-B_{mix} have a slight advantage (see Appendix 8.3).

Performance on Unseen VTU Tasks with TABLET We next evaluate whether TABLET's task diversity improves performance on unseen datasets. While fine-tuned models are expected to perform better on tasks seen during training, improved performance on held-out datasets would suggest that training on TABLET enhances the model's VTU capabilities. Tables 3 and 4 report our results on held-in and held-out tasks, respectively. We present comparisons between zero- and one-shot Qwen2.5-VL-7B⁴ and its fine-tuned instantiations trained on MMTab and different sizes of TABLET. For MMTab, all Wikipedia-sourced images are synthetic (61.8% of all tables in their dataset), whereas all TABLET versions use a mix of original (88%) and synthetic (12%) (See Appendix 7.2 for the distribution of image sources per task). For now, we focus solely on TABLET-L (a shorthand for TABLET-LARGE), which is our biggest dataset comprising 4M examples (we discuss smaller sizes in the next sections).

We observe that the model fine-tuned on TABLET-L performs overwhelmingly better on held-in *and* held-out datasets. It dominates in all QA benchmarks except HybridQA, where long-context profi-

⁴We should note that Qwen2.5-VL-7B is a competitive baseline that has undergone extensive pre-training and instruction-tuning, achieving top scores in many TU and Visual Document Understanding benchmarks.

ciency is likely reduced since its training set was excluded from fine-tuning. Notably, in tasks such as HiTab, FeTaQA, and TAT-QA, where TABLET includes fewer training examples than MMTab, the model still performs considerably better, suggesting benefits from transfer across tasks, original visualizations, or simply due to a higher-quality subset of examples. Finally, tasks that include hierarchical table structures such as ToTTo and HiTab show clear gains. Table cell hierarchy is not lost in synthetic tables, but it is often conveyed more clearly in their original visualizations. Our results are encouraging, particularly when considering that portions of the held-out datasets, including test sets, may have been encountered during petraining. Models trained on TABLET outperform 0/1-shot Qwen2.5-VL-7B on 12 out 14 (held-in and held-out) tasks.

Interestingly, our gains extend beyond unseen tasks. For instance, ToTTo, HiTab, TabFact, FeTaQA, and TAT-QA contain the *same* number of training examples across *all* TABLET variants, yet performance consistently improves as additional tasks are incorporated. This suggests that fine-tuning on a broader task set facilitates transfer learning, yielding improvements even on tasks where the model was already competitive.

Optimal TABLET Size Since training on 4M examples is resource-intensive, and tasks in TABLET-L exhibit substantial variation in size (see (b) pie chart in Table 4), we evaluate whether a smaller but more balanced dataset might perform comparably. We create TABLET-M (a short-hand for TABLET-MEDIUM) by capping each task at 140k examples. For benchmarks exceeding this cap (Column Type Annotation, Entity Linking, WikiBio), a random 140k subset is sampled and included as representative for that task. This results in a dataset with 1,117,217 training examples. We exclude HybridQA, InfoTabs, and Table Recognition from the training set to allow for held-out evaluation, leaving 1,031,082 examples in TABLET-M and 3,419,176 in the full dataset of our experiments. The capped distribution is shown in pie chart (b), Table 4 and further details on training sets are in Appendix 7.1.

The model trained on TABLET-M performs competitively and in some cases, even better than TABLET-L (see Table 4, rows in gray), offering a compelling trade-off between dataset size and model effectiveness. That said, the full dataset still yields slightly better performance overall and remains a valuable resource for future work that can benefit from large-scale, lossless training.

The Benefit of Training on Table Interpretation Tasks Table Interpretation tasks represent 54.2% of the training examples in TABLET-L, making them the largest category by volume and the second longest in instruction length, following HybridQA (see Appendix 8.1). While Deng et al. (2020) demonstrated the benefits of these tasks for textual TU, it is unclear whether this carries over to VTU. If these tasks prove unhelpful, removing them would save significant resources.

We therefore remove all Table Interpretation tasks from TABLET-M and fine-tune on this smaller version (we use TABLET-S as a shorthand for TABLET-SMALL) which excludes any tasks related to Column Type Annotation, Entity Linking, and Relation Extraction resulting in a dataset with a total of 690,467 training examples (67% of the example count in TABLET-M). As shown in Table 4, there are benefits to be gained from including these tasks. Models trained on TABLET-M outperform those trained on TABLET-S on six out of 8 held-in benchmarks (Table 3) and achieve comparable performance on the remaining two. Similar gains are observed in four out of five held-out tasks (Table 4). These findings demonstrate that the benefits observed in the textual domain (Deng et al., 2020) extend to multimodal VTU.

6 Conclusions

In this work, we introduced TABLET, a large-scale dataset for Visual Table Understanding that aggregates 4 million examples across 20 tasks and 2 million unique tables. Unlike prior resources, TABLET preserves original table visualizations whenever possible, provides both HTML and image representations, and maintains full traceability to the source datasets. Extensive experiments showed that (1) training on TABLET increases robustness to real-world tables compared to synthetic renderings; (2) models fine-tuned on TABLET consistently outperform those trained on related VTU datasets across held-in and held-out tasks; and (3) TABLET brings improvements on unseen benchmarks, which suggests that its diversity supports transfer across VTU tasks. We hope TABLET will foster further research on Visual Table Understanding, including the design of new tasks and the evaluation of models in settings that reflect real-world tabular data.

REFERENCES

- Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. PubHealthTab: A public health table-based dataset for evidence-based fact checking. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics:* NAACL 2022, pp. 1–16, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.1. URL https://aclanthology.org/2022.findings-naacl.1/.
- Iñigo Alonso, Eneko Agirre, and Mirella Lapata. PixT3: Pixel-based table-to-text generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6721–6736, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.364. URL https://aclanthology.org/2024.acl-long.364/.
- Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. Cont: Contrastive neural text generation. *arXiv preprint arXiv:2205.14690*, 2022.
- Chandra Sekhar Bhagavatula, Thanapon Noraset, and Doug Downey. Tabel: Entity linking in web tables. In Marcelo Arenas, Oscar Corcho, Elena Simperl, Markus Strohmaier, Mathieu d'Aquin, Kavitha Srinivas, Paul Groth, Michel Dumontier, Jeff Heflin, Krishnaprasad Thirunarayan, Krishnaprasad Thirunarayan, and Steffen Staab (eds.), *The Semantic Web ISWC 2015*, pp. 425–441, Cham, 2015. Springer International Publishing. ISBN 978-3-319-25007-6.
- 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. In *International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, April 2020a.
- Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *Findings of EMNLP* 2020, 2020b.
- 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. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1094–1110, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.78. URL https://aclanthology.org/2022.acl-long.78.
- Xiang Deng, Huan Sun, Alyssa Lees, You Wu, and Cong Yu. Turl: table understanding through representation learning. *Proceedings of the VLDB Endowment*, 14(3):307–319, 2020.
- Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web, 2023.
- Somraj Gautam, Abhishek Bhandari, and Gaurav Harit. TabComp: A dataset for visual table reading comprehension. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5773–5780, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.320. URL https://aclanthology.org/2025.findings-naacl.320/.
- Vivek Gupta, Maitrey Mehta, Pegah Nokhiz, and Vivek Srikumar. INFOTABS: Inference on tables as semi-structured data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2309–2324, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.210. URL https://aclanthology.org/2020.acl-main.210/.
- Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. TaPas: Weakly supervised table parsing via pre-training. In Dan Jurafsky, Joyce Chai,

Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4320–4333, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.398. URL https://aclanthology.org/2020.acl-main.398.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3096–3120, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. findings-emnlp.175. URL https://aclanthology.org/2024.findings-emnlp.175/.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5817–5834, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.291. URL https://aclanthology.org/2025.acl-long.291/.

- Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. Tabmcq: A dataset of general knowledge tables and multiple-choice questions, 2016. URL https://arxiv.org/abs/1602.03960.
- Jun-Peng Jiang, Yu Xia, Hai-Long Sun, Shiyin Lu, Qing-Guo Chen, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Multimodal tabular reasoning with privileged structured information, 2025. URL https://arxiv.org/abs/2506.04088.
- Yannis Katsis, Saneem Chemmengath, Vishwajeet Kumar, Samarth Bharadwaj, Mustafa Canim, Michael Glass, Alfio Gliozzo, Feifei Pan, Jaydeep Sen, Karthik Sankaranarayanan, and Soumen Chakrabarti. AIT-QA: Question answering dataset over complex tables in the airline industry. In Anastassia Loukina, Rashmi Gangadharaiah, and Bonan Min (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pp. 305–314, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-industry.34. URL https://aclanthology.org/2022.naacl-industry.34/.
- Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyoong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision (ECCV)*, 2022.
- Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains, 2024. URL https://arxiv.org/abs/2404.19205.
- Rémi Lebret, David Grangier, and Michael Auli. Neural text generation from structured data with application to the biography domain. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1203–1213, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1128. URL https://aclanthology.org/D16-1128/.
- Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding, 2023. URL https://arxiv.org/abs/2210.03347.
- Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1966.
- Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. TableBank: Table benchmark for image-based table detection and recognition. In Nicoletta Calzolari, Frédéric

Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 1918–1925, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.236/.

- Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-gpt: Table-tuned gpt for diverse table tasks, 2023. URL https://arxiv.org/abs/2310.09263.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- Qian Liu, Bei Chen, Jiaqi Guo, Morteza Ziyadi, Zeqi Lin, Weizhu Chen, and Jian-Guang Lou. TAPEX: Table pre-training via learning a neural SQL executor. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=050443AsCP.
- Zhenghao Liu, Haolan Wang, Xinze Li, Qiushi Xiong, Xiaocui Yang, Yu Gu, Yukun Yan, Qi Shi, Fangfang Li, Ge Yu, and Maosong Sun. Hippo: Enhancing the table understanding capability of large language models through hybrid-modal preference optimization, 2025. URL https://arxiv.org/abs/2502.17315.
- 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. In *International Conference on Learning Representations (ICLR)*, 2023.
- Yadong Lu, Jianwei Yang, Yelong Shen, and Ahmed Awadallah. Omniparser for pure vision based gui agent, 2024. URL https://arxiv.org/abs/2408.00203.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.
- Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, Mutethia Mutuma, Ben Rosand, Isabel Trindade, Renusree Bandaru, Jacob Cunningham, Caiming Xiong, Dragomir Radev, and Dragomir Radev. FeTaQA: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022. doi: 10.1162/tacl_a_00446. URL https://aclanthology.org/2022.tacl-1.3/.
- National Center for Science and Engineering Statistics. Science and engineering indicators 2019, 2019. URL https://www.nsf.gov/statistics/2019/nsf19319/. Accessed: 2024-10-28.
- Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.89. URL https://aclanthology.org/2020.emnlp-main.89.
- Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In Chengqing Zong and Michael Strube (eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1142. URL https://aclanthology.org/P15-1142.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li,

649

650

651

652

653

654

655

656

657

658

659

661

662

663

665

667

668

669

670

671 672

673

674

675

676

677

678

679

680

684

685

686

687

688

689

690

691

692

693

696

697

Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hui Shi, Yusheng Xie, Luis Goncalves, Sicun Gao, and Jishen Zhao. Wikidt: Visual-based table recognition and question answering dataset. *Proceedings of the 18th International Conference on Document Analysis and Recognition (IC-DAR)*, 2024. URL https://www.amazon.science/publications/wikidt-visual-based-table-recognition-and-question-answering-dataset.

Statistics Canada. Statistics canada - the national statistical office of canada, 2024. URL https://www150.statcan.gc.ca. Accessed: 2024-10-28.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Gang Chen, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. Tablegpt2: A large multimodal model with tabular data integration, 2024. URL https://arxiv.org/abs/2411.02059.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam

Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

- Prasham Yatinkumar Titiya, Jainil Trivedi, Chitta Baral, and Vivek Gupta. Mmtbench: A unified benchmark for complex multimodal table reasoning, 2025. URL https://arxiv.org/abs/2505.21771.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.
- Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Guohai Xu, Chenliang Li, Junfeng Tian, Qi Qian, Ji Zhang, Qin Jin, Liang He, Xin Lin, and Fei Huang. UReader: Universal OCR-free visually-situated language understanding with multimodal large language model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2841–2858, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.187. URL https://aclanthology.org/2023.findings-emnlp.187.
- Jing Zhang, Zhi Chen, Zhiyuan Liu, and Maosong Sun. Graph neural networks for table structure recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3610–3620. Association for Computational Linguistics, 2020.
- Tianshu Zhang, Xiang Yue, Yifei Li, and Huan Sun. TableLlama: Towards open large generalist models for tables. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6024–6044, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024. naacl-long.335. URL https://aclanthology.org/2024.naacl-long.335.
- Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding, 2023.
- Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Houqiang Li, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *arXiv preprint arXiv:2406.01326*, 2024.
- Boyuan Zheng, Boyu Gou, Jihyung Kil, Huan Sun, and Yu Su. GPT-4v(ision) is a generalist web agent, if grounded. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=piecKJ2DlB.
- Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9102–9124, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.493. URL https://aclanthology.org/2024.acl-long.493/.
- Xu Zhong, Elaheh ShafieiBavani, and Richard Zanibbi. Image-based table recognition: data, model, and evaluation. *Proceedings of the 16th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 847–854, 2020.
- Wei Zhou, Mohsen Mesgar, Heike Adel, and Annemarie Friedrich. Texts or images? a fine-grained analysis on the effectiveness of input representations and models for table question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), Findings of the Association for Computational Linguistics: ACL 2025, pp. 2307–2318, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.117. URL https://aclanthology.org/2025.findings-acl.117/.

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. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.254. URL https://aclanthology.org/2021.acl-long.254/.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

7 APPENDIX

7.1 TABLET TRAINING SET STATISTICS

While the full released version of TABLET includes all available data, our experiments used different partitions. Specifically, to allow for held-out evaluation, we excluded HybridQA, InfoTabs, and Table Recognition from the training set; in addition, we created three training sets varying in size (TABLET-L, TABLET-M, TABLET-S) to explore various research questions. The statistics for these training sets are reported in Table 1. The development and test sets remain identical to the ones mentioned in the paper.

	Examples	Tasks	Original Images
TABLET-S	690,467	14	521,032 (75.5%)
Tablet-M	1,031,082	17	812,748 (78.8%)
TABLET-L	3,419,176	17	2,795,485 (81.8%)

Table 5: Training set statistics for TABLET-L, TABLET-M, and TABLET-S used to fine-tune Qwenn2.5-VL 7B in our experiments. We report the number of examples (Examples), the tasks used (Tasks), and the number of examples with original table visualizations.

7.2 IMAGE SOURCE PER TASK

Figure 3 shows the distribution of images per task included in TABLET. Images are sourced from Wikipedia, TabMWP, PubTabNet or are syntehtically rendered from information in the seed dataset.

7.3 Dataset Crawling Dates

We determined data collection dates for each dataset through various sources and methodologies. TURL originates from the TabEl dataset (Bhagavatula et al., 2015), which was crawled from the November 2013 English Wikipedia dump. For ToTTo, we obtained the collection date through direct correspondence with Ankur Parikh. TabFact's timeline was established based on email correspondence with Wenhu Chen and the initial arXiv publication date.

For HybridQA, we conducted a comprehensive analysis of all dates present within the dataset tables and observed a significant decrease in data frequency from February 2020 onward, indicating the collection cutoff point. InfoTabs' crawling year is documented at the bottom of their GitHub repository page, though the specific day represents our estimation. WikiBio's GitHub documentation indicates that their tables are sourced from the Common Crawl snapshot enwiki-20150901.

Dataset	Date
TURL	2013-11-01
ToTTo	2019-03-01
TabFact	2019-06-30
HybridQA	2020-01-31
Infotabs	2019-10-10
WikiBio	2015-09-01

Table 6: Data collection timeline for benchmarks included in TABLET.

7.4 Dataset example

For reference, we show an example from TABLET for the ToTTo table-to-text task. Note that examples for every other task follow the same format. We provide the instructions used in our experiments, together with the corresponding example metadata, and links to the corresponding example and table from the source dataset.

7.5 DEFINITION OF TASKS REPRESENTED IN TABLET

As mentioned in Section 3.3, TABLET includes 20 TU tasks. We provide more detailed descriptions for each below.

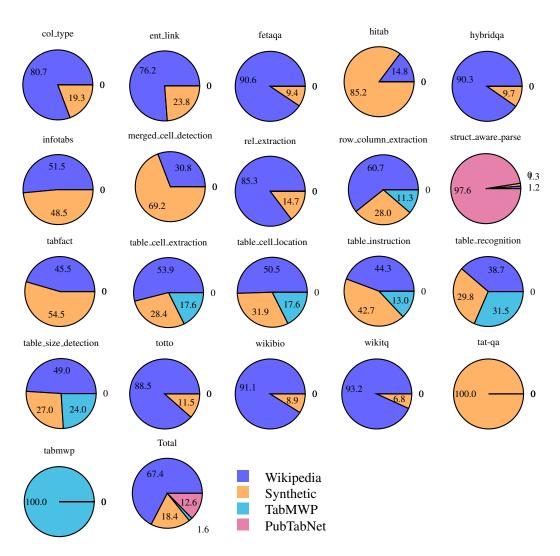


Figure 2: Image source distribution in TABLET, broken down by task. That is, source of the image referred by each example in each task. While the distribution resembles that of the unique image pool, it is computed at the example level (e.g., if the same Wikipedia image appears in two examples of a task, it is counted twice). Wikipedia are original visualizations form Wikipedia. Seed render are synthetic images rendered form information in the seed dataset.

Column Type Annotation In this task, the model is required to identify the data type of the values in a highlighted table column. The model selects from a set of 255 randomly shuffled candidate data types. There can be multiple correct types per column. The column is visually highlighted to ensure that the model attends to the table itself rather than relying solely on the textual instruction to infer the answer. Examples for this task were obtained from TURL. Our dataset includes 602,406/13,188/12,802 (train/dev/test) instructions for this task. An example of this task is shown in Figure 17.

Entity Linking For this task, the model is given a table with a visually highlighted cell along with a set of up to 100 candidate entities and descriptions and must identify which entity and description corresponds to the entity in the selected cell. This task was aggregated from the TURL dataset. Our dataset includes 1,236,128/74,282/213,494 (train/dev/test) instructions for this task. An example of this task is shown in Figure 5.

```
918
                        "example_id": "e0af904e6617d0ab211ae76e62f7308c",
"html_path": "html/highlighted/ToTTo/totto/train/254d984c233855f5be6814c6d1388c92_e0af904e6617d0ab211ae76e62f7308c.html",
"img_path": "img/highlighted/ToTTo/totto/train/254d984c233855f5be6814c6d1388c92_e0af904e6617d0ab211ae76e62f7308c.png",
919
920
                        "html_raw_path": "html/raw/TOTTo/totto/254d984c233855f5be6814c6d1388c92.html",
"img_raw_path": "img/raw/TOTTo/totto/254d984c233855f5be6814c6d1388c92.png",
"img_source": "wikipedia",
"input": "Taking into account the table information on 'List of 8/9 PM telenovelas of Rede Globo', section '2000s', produce a
921
922
                   ingle-sentence summary focused on the highlighted cells and format it within the following JSON object {\"answer\": \"YOUR
923
                         output": "{\"answer\": \"A Favorita is the telenovela aired in the 9 pm timeslot.\"}",
924
                        "metadata": {"table_page_title": [46, 86], "table_section_title": [98, 103], "final_sentence": [12, 68]}, 
"seed_table_id": "1762238357686640028:train:0",
925
                        "split": "train"
                        "src_example_ids": {
    "ToTTo": "1762238357686640028:train:0"
926
927
                         table_id": "254d984c233855f5be6814c6d1388c92"
928
                        "table_page_title": "List of 8/9 PM telenovelas of Rede Globo", "table_section_title": "2000s", "table_seed_dataset": "ToTTo",
929
930
                         "table_variant": "highlighted",
                        "task": "totto",
931
                        "table wiki page id": "45544626".
                        "table_wiki_old_id": 876845524,
932
```

Figure 3: Example of a TABLET example for the ToTTo table-to-text task.

Relation Extraction This task requires the model to select appropriate relations between two visually highlighted columns of a table from a set of candidates. This task is derived from the TURL dataset. Our dataset includes 60,615/2,145/2,030 (train/dev/test) instructions for this task. An example of this task is shown in Figure 7.

Structure Aware Parsing In this TSR task, the model needs to parse the table into markdown format. This task comes from Docstruct4M, using tables from PubTabNet, TabFact, and WikiTable-Questions. Our dataset includes 513,482/9,115/1,102 (train/dev/test) instructions for this task. An example of this task is shown in Figure 9.

Free-form Table Question Answering (FeTaQA) In this task, the model generates free-form answers to questions about Wikipedia tables, often requiring integration of information from discontinuous sections of the table. Unlike datasets with shorter text spans, FeTaQA emphasizes higher-level understanding through long-form answers. Examples for this task come from FeTaQA dataset. Our dataset contains 3,006/577/1,079 (train/dev/test) instructions for this task. An example of this task is shown in Figure 6.

Hierarchical Table QA (HiTabQA) This question-answering task involves hierarchical tables (with different headers across the table) and sometimes includes numerical reasoning, such as sums, averages, maximum, minimum, and counting, among others. Examples for this task were obtained from HiTabQA, with tables from ToTTo, StatCan, and NSF seed datasets. Our dataset contains 7,417/1,670/1,584 (train/dev/test) examples. An example of this task is shown in Figure 8.

Table Fact Verification (TabFact, Infotabs) Also known as Table Entailment, this task involves classifying statements as supported or refuted based on table content. Examples for these two tasks come from TabFact and Infotabs seed datasets. Our dataset includes 87,717/12,389 / 12,326 (train / dev / test) TabFact examples and 16,538 / 1,800 / 5,400 (train / dev / test) Infotabs examples. An example of one of these tasks is shown in Figure 11.

Table Numerical Reasoning (TabMWP, TAT-QA) Given a table and a mathematical question, the model must answer using mathematical reasoning over table values. Instructions were based on examples from TabMWP and TAT-QA. Our dataset includes 23,059 / 7,686 / 7,686 for TABMWP and 2,201 / 278 / 277 (train / dev / test) examples. An example of one of these tasks is shown in Figure 10.

Table-to-Text (ToTTo) In this task, the model needs to generate a description based on the visually highlighted cells in a given table. Instructions were generated based on examples from ToTTo. Not all examples from the original dataset were retrieved, as we could not trace all highlighted cells from

the dataset to the retrieved table for 8.1% of the examples. These examples were discarded to avoid adding noise to the dataset. Our dataset contains 110,934/7,077/7,084 (train/dev/test) examples. An example of this task is shown in Figure 12.

Hybrid QA (**HybridQA**) This multi-hop question-answering task requires integrating structured table data with unstructured hyperlinked passages. Given a Wikipedia table and texts linked to the table's entities, the model must answer multi-hop questions by reasoning across both modalities. Instructions were generated using HybridQA's examples, resulting in a dataset containing 62,670/3,466/3,463 (train/dev/test) examples. An example of this task is shown in Figure 14.

Table QA (WikiTableQuestions) Given a Wikipedia table and a question, the model must answer based on the table's content. For this task, WikiTableQA's examples are phrased as instructions in our dataset. Our dataset includes 14,152/3,537/4,344 (train/dev/test) instructions. An example of this task is shown in Figure 16.

Wikipedia Biography Generation (WikiBio) Given a Wikipedia infobox of an entity, the model is prompted to generate a concise biography of this entity using the information in the infobox. This task's examples are aggregated from WikiBio. Our dataset includes 582,659/72,831/72,831 (train/dev/test) instructions. An example of this task is shown in Figure 13.

MMTab's Structure Understanding tasks TABLET includes all tasks from MMTab that are meant to instill table structure understanding in the model. These include: Merged Cell Detection, Row & Column Extraction, Table Cell Extraction, Table Cell Location, Table Recognition, Table Size Detection, and instruction following pre-training tasks. We refer to Zheng et al. (2024b)'s work for a description of these tasks. Instructions for these tasks are directly aggregated from MMTab and tables come from the following seed datasets: InfoTabs, NSF, StatCan, TabMWP, TAT-QA, TabFact, ToTTo, WikiBIO, WikiTableQuestions. TABLET maintains the same instructions as in MMTab but uses our visualizations for the table images. As instructions originate from MMTab, and no dev set is provided in this dataset, TABLET does not include any example for these tasks in the development set. Our dataset includes the following examples: Merged Cell Detection (7,500/0/950), Row & Column Extraction (7,721/0/957), Table Cell Extraction (7,727/0/966), Table Cell Location (7,708/0/956), Table Recognition (6,927/0/843), Table Size Detection (7,800/0/950), and instruction following pre-training tasks (136,944/0/0). An example of one of these tasks is shown in Figure 17.

7.6 MMTAB VS TABLET

Table 7 provides a detailed comparison between MMTab and TABLET across tasks and example counts (in training, development and test sets).

8 SUPERVISED FINE-TUNING DETAILS

All supervised fine-tuning (SFT) experiments were carried out with the same hyperparameters across models; the only varying factor was the dataset used for training. We fine-tuned **Qwen2.5-VL-7B-Instruct** using their official implementation.⁵ Our code will be released alongside the dataset in their project repository.

Hyperparameters All runs used the following common configuration:

- Training setup: DeepSpeed ZeRO-3, bf16 precision
- Epochs: 3
- Batch size: 2 (per device), gradient accumulation steps: 4
- Optimizer: AdamW, learning rate: 2e-7, weight decay: 0.01
- Scheduler: cosine decay with warmup ratio 0.03
- Gradient clipping: 1.0

 $^{^{5} \}texttt{https://github.com/QwenLM/Qwen2.5-VL/tree/main/qwen-vl-finetune}$

	T	rain	D	ev	Te	est
Task	MMTab	TABLET	MMTab	TABLET	MMTab	TABLET
wikibio	4,994	582,659	0	72,831	1,000	72,831
wikitq	17,689	14,152	0	3,537	4,344	4,344
totto	15,000	110,934	0	7,077	7,700	7,084
tabmwp	30,745	23,059	0	7,686	7,686	7,686
tabfact	31,321	87,717	0	12,389	6,845	12,326
hitab	11,941	7,417	0	1,670	3,160	1,584
infotabs	18,338	16,538	0	1,800	5,400	5,400
fetaqa	8,327	3,006	0	577	2,003	1,079
tat-qa	5,920	2,201	0	278	772	277
table_instruction	37,204	136,944	0	0	0	0
table_cell_extraction	8,000	7,727	0	0	1,000	966
table_cell_location	8,000	7,708	0	0	1,000	956
table_size_detection	8,000	7,800	0	0	1,000	950
merged_cell_detection	8,000	7,500	0	0	1,000	950
row_column_extraction	8,000	7,721	0	0	1,000	957
table_recognition	8,000	6,927	0	0	1,000	912
rotowire	3,400	0	0	0	334	0
col_type	0	602,406	0	13,188	0	12,802
ent_link	0	1,236,128	0	74,282	0	213,494
rel_extraction	0	60,615	0	2,145	0	2,030
hybridqa	0	62,670	0	3,466	0	3,463
struct_aware_parse	0	513,482	0	9,115	0	1,102
OOD	0	0	0	0	1,250	0
TabMCQ	0	0	0	0	1,029	0
AIT-QA	0	0	0	0	511	0
PubHealthTab	0	0	0	0	1,942	0
All	232,879	3,505,311	0	210,041	49,976	351,193

Table 7: Comparison of MMTab and TABLET: tasks and examples across training, development, and test splits.

- Sequence length: 8192 tokens
- Vision input size: max_pixels = 50,176, min_pixels = 784
- Other: data_flatten = False, data_packing = False, tune_mm_vision = False, tune_mm_mlp = True, tune_mm_llm = True

Compute Usage Table 8 reports the total GPU hours consumed by each experiment. All runs were conducted on clusters equipped with NVIDIA A100 GPUs.

8.1 Instruction Sequence Length

Figure 4 shows how instruction length (measured in terms of tokens) varies across tasks in TABLET. As can be seen, HybridQA has the longest instructions, followed by Column Type, Entity Linking, and Relation Extraction.

8.2 Performance Degradation: Original vs Synthetic Images

Let each task t belong to one of the metric families \mathcal{B} (BLEU), \mathcal{A} (accuracy), or \mathcal{F} (F1). For a given model m, we compute the raw difference:

Dataset	Setup	GPU Hours
TABLET-B _{org}	$15h \times 16$ GPUs	240
TABLET-B _{synth}	$15h \times 16 \text{ GPUs}$	240
TABLET-B _{mix}	$21h \times 16 \text{ GPUs}$	336
MMTab	$21h \times 16 \text{ GPUs}$	336
Tablet-S	$28h \times 32 \text{ GPUs}$	896
Tablet-M	$40h \times 32 \text{ GPUs}$	1280
TABLET-L	$125h \times 32 \text{ GPUs}$	4000
Inference (TABLET-B _{org})	21h × 8 models	168
Inference (TABLET-B _{synth})	$21h \times 8$ models	168
Inference (TABLET-B _{mix})	$24h \times 8$ models	192
Inference (MMTab)	$24h \times 8$ models	192
Inference (TABLET- B_L)	$34h \times 8 \text{ models}$	272

Table 8: GPU hours for supervised fine-tuning and prediction runs. For the predictions, the 8 models are the 7 SFT models and the baseline model.

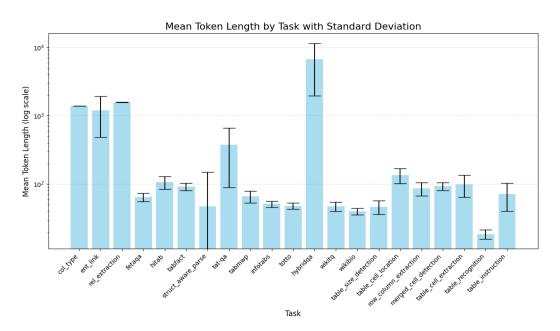


Figure 4: Instruction length distribution across tasks in TABLET.

$$\Delta_{m,t} = \operatorname{Score}_{m,t}^{\operatorname{org}} - \operatorname{Score}_{m,t}^{\operatorname{synth}}$$

Negative values indicate that the model performs worse when evaluated on the original visualizations compared to the synthetic ones. Because the metrics differ in scale, we normalize each by the corresponding synthetic score to report results in terms of percentage change:

$$\delta_{m,t} = \frac{\Delta_{m,t}}{\text{Score}_{m,t}^{\text{synth}}} \times 100.$$

Next, for each metric family $F \in \mathcal{B}, \mathcal{A}, \mathcal{F}$, we compute the family mean (M). Finally, the Degradation Score for model m is defined as the unweighted average of the family means:

$$DegScore = \frac{1}{3} \Big(M_{m,\mathcal{B}} + M_{m,\mathcal{A}} + M_{m,\mathcal{F}} \Big).$$

This guarantees that BLEU, accuracy, and F1 contribute equally, regardless of how many tasks are included in each family. Our results are summarized in Table 9 below where we show the Degration Score per dataset and overall. Results are reported only for tasks with synthetic *and* original images. InfoTabs is held-out, all other tasks are held-in. As can be seen in the table, the 0-shot model performs worse when evaluated on original visualizations, followed by Qwen2.5-VL-7B fine-tuned on TABLET-B_{synth}. Models which have seen original images during fine-tuning perform best (see TABLET-B_{org} and TABLET-B_{mix}.

Model	WikiBio '	ГоТТо	WikiTQ	HiTab	TabFact	FeTaqa	Infotabs	(BLEU)	(acc)	(F1)	DegScore
0-Shot	+0.5	-1.9	-6.9	-8.6	-11.0	-0.1	-4.6	-1.5	-22.5	-8.6	-28.90
TABLET-B _{synt}	h -6.5	+2.0	-1.9	-17.5	+2.4	-0.7	-0.6	-5.2	-0.1	-17.5	-22.35
TABLET-Borg	+1.9	+2.2	-2.3	-10.4	+1.6	+0.0	-2.7	+4.1	-3.4	-10.4	-6.63
TABLET- B_{mix}	+1.7	+1.6	-2.4	-11.6	+1.8	-0.1	-4.8	+3.2	-5.4	-11.6	-7.87

Table 9: Comparison of **Qwen2.5-VL-7B-Instruct** models fine-tuned on different TABLET-B variants and in zero-shot mode; models are evaluated on original (lossless) table visualizations and synthetic ones across held-in benchmarks. Totals are summed separately for BLEU, accuracy (acc), and F1; the average percentage-point change (DegScore) is a unit-less normalized aggregate (more negative is worse).

8.3 TABLET-B HELD-OUT EVALUATION RESULTS

In Table 10, we present results on held-out portions of TABLET-B, with Qwen2.5-VL-7B fine-tuned on original table images (TABLET- B_{org}), synthetic ones (TABLET- B_{synth}), and a mixture (TABLET- B_{mix}).

	Infotabs		TabMCQ	AIT-QA	PubHealthTab
Model	mix org		org	org	org
TABLET-B _{synth}	52.4	51.8	87.8	64.4	57.6
TABLET-B _{org}	51.4	48.7	87.7	63.8	55.0
$TABLET\text{-}B_{mix}$	60.3	55.5	88.3	59.5	61.6

Table 10: Comparison of **Qwen2.5-VL-7B** fine-tuned on different MMTab variants evaluated on purely *lossless* (org) table visualizations and a *mix* of both across held-out benchmarks. Results for TabMWP, AIT-QA, and PubHealthTab do not include *mix* because the tables in these tasks are original visualizations. We report accuracy for all benchmarks.

8.4 OPEN-WEIGHT MODEL EVALUATION ON TABLET

Finally, in Tables 11 and 12, we report results on TABLET for open-weight VLMs other than Qwen2.5-VL-7B, in the zero-shot setting.

	WikiBio	ToTTo	WikiTQ	TabMWP	HiTab	TabFact	FeTaQA	TAT-QA
Table-LLaVA 7B (Zheng et al., 2024b)	5.7	16.3	17.7	47.5	10.6	53.6	7.5	1.8
Qwen2.5-VL 7B (Qwen et al., 2025)	2.5	9.1	53.4	59.1	31.2	73.9	7.0	6.9
DocOwl2 8B (Hu et al., 2025)	1.8	1.2	18.9	0.5	9.5	54.3	9.6	2.9
InternVL3 8B (Zhu et al., 2025)	4.6	11.9	45.3	83.9	40.8	64.3	3.8	13.0
InternVL3 14B (Zhu et al., 2025)	6.3	12.4	51.6	86.8	52.0	75.1	6.6	11.9
Gemma-3 12B (Team et al., 2025)	5.1	5.4	39.1	72.3	24.2	69.5	20.3	2.5

Table 11: Open-weight VLMs evaluated on TABLET held-in datasets (zero-shot setting).

	Infotabs	TabMCQ	AIT-QA	PubHealthTab	HybridQA
Table-LLaVA 7B (Zheng et al., 2024b)	61.5	59.1	4.7	50.2	_
Qwen2.5-VL 7B (Qwen et al., 2025)	64.5	84.4	51.7	64.7	34.2
DocOwl2 8B Hu et al. (2025)	15.5	63.0	36.8	14.6	_
InternVL3 8B (Zhu et al., 2025)	59.6	88.5	52.4	64.2	33.1
InternVL3 14B (Zhu et al., 2025)	67.4	88.8	52.3	74.9	46.7
Gemma-3 12B (Team et al., 2025)	65.5	88.9	48.7	64.7	35.2

Table 12: Open-weight VLMs evaluated on TABLET held-out datasets (zero-shot setting). Results for HybridQA are not reported for Table-LLaVA 7B and DocOwl2 8B due to their VRAM requirements, which do not scale well to the long contexts needed for our dataset.

Table:

Place	Athlete	Time	Qual.
1	Mel Brock (CAN)	1:57.0	QS
2	Ted Meredith (USA)		QS
3	■ John Victor (RSA)		
_	Alan Patterson (GBR)	DNF	

Instruction:

```
For the Wikipedia table from the

article 'Athletics at the 1912

Summer Olympics Men's 800 metres'

see 'Heats', select the proper

entity that matches the highlighted

table value given the following

possible entities (<name /

description / type>): <Alan

Patterson / Wikipedia

disambiguation page / None>, <Alan

Patterson / British athlete /

owl#Thing>, <Alan Patterson / UK MP

/ Person>, [...]. Return only the

identifier or name of the chosen

entity as JSON: {"answer": "YOUR

ANSWER"}.
```

Expected output:

```
{"answer": "<Alan Patterson / British

→ athlete / owl#Thing>"}
```

Figure 5: Example for Entity Linking task based on highlighted table cell.

9 LIMITATIONS

While TABLET is the largest resource for Visual Table Understanding to date, it has several limitations. Firstly, some original visualizations could not be retrieved due to missing or changed Wikipedia pages, and some embedded resources (e.g., images) were inaccessible. Secondly, we did not conduct a full ablation of the contribution of each task due to computational constraints, focusing instead on broader questions such as task balancing and dataset inclusion. Thirdly, most seed datasets are from Wikipedia, which, despite differing in visual format, are common in pretraining corpora, raising the possibility of data contamination. Finally, our fine-tuning focused on a single model (Qwen2.5-VL 7B), with zero-shot evaluation on others; future work should explore a broader range of architectures, especially those with reasoning or self-reflection capabilities.

Year +	Competition +	Venue +	Position +	Event +	Notes +				
	Representing Soviet Union								
1976	Olympic Games	Montreal, Canada	3rd	4×100 m relay					
1978	European Indoor Championships	Milan, Italy	1st	60 metres					

Instruction:

```
According to the table titled Nikolay Kolesnikov

(sprinter), section Achievements, write a short

sentence answer to the question: How did Nikolay

Kolesnikov do at the 1976 Olympics and at the 60

metres at the 1978 European Indoor Championships?

Output only the correct answer as {"answer": "YOUR

ANSWER"}.
```

Expected output:

```
{"answer": "Nikolay Kolesnikov won a bronze medal at \hookrightarrow the 1976 Olympics and won the 60 metres at the \hookrightarrow 1978 European Indoor Championships."}
```

Figure 6: Example for Free-form Table Question Answering task based on highlighted table cells (FeTaQA).

Table:

Nationality	Player	Ranking*	Seeding
 ISR	Dudi Sela	56	1
TPE	Lu Yen-hsun	62	2
■ USA	Bobby Reynolds	93	3
GER	Michael Berrer	111	4
BRA	Thiago Alves	120	5
GER	Benjamin Becker	126	6
■ FRA	Nicolas Mahut	137	7
RUS	Michail Elgin	138	8

Instruction:

```
For the table found in the article '2009 israel open' (section 'seeds') on Wikipedia, determine which relation holds between the two highlighted columns among these relation candidates:

base.wikipedia_infobox.video_game.developer, organization.organization.headquarters, award.award_nominated_work.award_nominations, award.award_nomination.award_nominee, [...].

Provide the correct relation as JSON:

"answer": ["RELATION"]} - list multiple relations comma-separated if necessary.
```

```
{"answer":

→ ["people.person.nationality"]}
```

Figure 7: Example for Relation Extraction task based on highlighted table columns.

Practice summary

Session	Dav					
Session	No No		Driver	Team	Car	Time
Practice 1	Friday	9	Marid Reynolds	Erebus Motorsport	Holden VF Commodore	1:12.8697
Practice 2	Friday	88	Marie Whincup	Triple Eight Race Engineering	Holden VF Commodore	1:12.5530

Instruction:

```
Answer the given question as concisely as possible

→ using information from the table. Answers must be

→ within a JSON object e.g. {"answer": ["YOUR

→ ANSWER"]}. List multiple answers with commas

→ inside the list object. Table caption: 'practice'.

→ Relying on the table, answer the following: What

→ was the time of david reynolds in practice 1?
```

Expected output:

```
{"answer": ["1:12.8697"]}
```

Figure 8: Example for Hierarchical Table QA task (HiTabQA).

Table:

Diagnosis	Cases	Indeterminate	Evaluable
Follicular	34	6/34 (17.6%)	28/34 (82.4%)
Mantie cell	7	0/7 (0%)	7/7 (100%)
Marginal Zone	11	1/11 (9%)	10/11 (91%)
Lymphoplasmacytic	6	0/6 (0%)	6/6 (100%)
Diffuse Large B Cell	15	2/15 (13.3%)	13/15 (86.7%)
CLL/SLL	4	0/4 (0%)	4/4 (100%)
Small B Cell NHL, NOS	2	0/2 (0%)	2/2 (100%)
Total	79	9/79 (11.3%)	70/79 (88.6%)

Instruction:

Convert the illustration to Markdown syntax.

Expected output:

Figure 9: Example for Structure Aware Parsing task.

Table: Plants per garden

Stem	Leaf
3	33355
4	6
5	4578
6	78
7	2379
8	689

Instruction:

```
The members of the local garden club tallied the

→ number of plants in each person's garden. How many

→ gardens have at least 47 plants?

Solve the above problem according to the table called

→ 'Plants per garden'. Include your reasoning, after

→ which finish with the JSON {"answer": "YOUR

→ ANSWER"}.
```

```
Find the row with stem 4. Count all the leaves greater \hookrightarrow than or equal to 7.

Count all the leaves in the rows with stems 5, 6, 7, \hookrightarrow and 8.

You counted 13 leaves, which are blue in the \hookrightarrow stem-and-leaf plots above. 13 gardens have at \hookrightarrow least 47 plants. Therefore {"answer": "13"}
```

Figure 10: Example for Table Numerical Reasoning task (TabMWP).

1350 1351

1352

1353

1354 1355

1356

1357

1358

1359

1360

1361

1362

1363

1364

1365

1366 1367 1368

1374

1375

1376

1377

1378

1379 1380

1381

1383

1384 1385 1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1398

1399

1400

1401 1402

1403



Instruction:

Consulting the table, state whether the hypothesis is → entailed, neutral, or refuted. Give your final → answer in JSON format {"answer": "YOUR ANSWER"}.

Hypothesis: Charles Francis Jenkins has been awarded \hookrightarrow more than one medal.

Expected output:

```
{"answer": "entailed"}
```

Figure 11: Example for Table Fact Verification task (Infotabs).

Table:

Date	Time		Score		Set 1	Set 2	Set 3	Set 4	Set 5	Total	Report
20 Jul	19:30	Slovakia 👊	3-1	Netherlands	25- 21	25- 18	20- 25	25- 23		95-87	Report

Instruction:

```
Consulting the following table concerning '2008 Men's

→ European Volleyball League' in the 'Final'

→ section, give a one-line statement limited to the

→ highlighted cells and return it as JSON:

→ {"answer": "YOUR ANSWER"}.
```

Expected output:

```
{"answer": "The 2008 Men's European Volleyball League \hookrightarrow was won by Slovakia, defeating the Netherlands by \hookrightarrow 3{1 in the finals."}
```

Figure 12: Example for Table-to-Text task based on highlighted table cells (ToTTo).

Table:

```
Piet van Zyl
Birth name Piet van Zyl

Date of birth May 14, 1979 (age 46)<sup>[1]</sup>
Place of birth ≥ Worcester, South
              1.81 m (5 ft 11 in)[1]
Weight
                 95 kg (209 lb)[1
       Rugby union career
Position(s) Centre
               Senior career
Correct as of 21 May 2013
         Provincial / State sides

        Years
        Team
        Apps (Points)

        2006-08
        Boland Cavaliers
        28
        30<sup>[2]</sup>

Correct as of 21 May 2013
        Super Rugby
Team
                                  Apps (Points)
2008-09
Correct as of 21 May 2013
                                 Apps (Points)
Correct as of 9 September 2007
```

Instruction:

```
Drawing upon the provided infobox of piet van zyl

→ -lrb- namibian rugby union player -rrb-, write a

→ third-person, encyclopedic biography and format it

→ within the following JSON object {"answer": "YOUR

→ BIOGRAPHY"}.
```

```
{"answer": "Piet van zyl (born 14 may 1979) is a

→ namibian rugby union player who captained the

→ boland cavaliers in south africa at provincial

→ level, and played for the at international level.

→ Van zyl was in the namibian squad for the 2007

→ world cup, and scored a try in his nation 's first

→ match in the competition, in a game against. Van

→ zyl plays as a centre. Van zyl made his debut in

→ august 2007 in a friendly match against."}
```

Figure 13: Example for Wikipedia Biography Generation task (WikiBio).

Table:	Rank •	Gymnasts +	Country +	Point •
	1	Attila Katus, Tamas Katus, Romeo Szentgyorgyi	Hungary	16.55
	2	Dorel Mois, Claudiu Moldovan, Claudiu Varlam	■■ Romania	16.25
	3	Maria Holmgren, Helene Nilsson, Kim Wickman	Sweden	15.87
	4	Stanislav Marchenkov, Vadim Mikhailov, Denis Belikov	Russia	15.55
	5	Grégory Alcan, Xavier Julien, Olivier Salvan	■ France	15.00
	6	Won-Sil Choi, Hyun-Sung Ki, Kwang-Soo Park	South Korea	14.95
	7	Marie-Catherine Boesa, Jana Heinze, Sandra Schlueter	Germany	14.835
	8	Yumi Kobayashi, Kumi Sato, Hiroko Watabe	Japan	13.758
	9	Giacomo Piccoli, Giovanna Lecis, Marco Bisciaio	■ Italy	13.044

Instruction:

1404 1405

1416

1417

1418

1419

1420

1422

1424

1425

1426

1427

1428

1429

1430

1431

1432

1433

1434

1435

1437

1438 1439

1441 1442 1443

1444

1445

1446

1447

1448

1449

1450

1451

1452

1453

1454

1455

1456 1457

```
Taking into account the accompanying table as well as
\hookrightarrow the context snippets provided at the end.
Answer the following question: What is the official
\hookrightarrow name of the country that finished with the fifth
    most points at the 1998 Aerobic Gymnastics World
    Championships ? Respond with the correct answer
    (omit explanations) in JSON format as {"answer":
    "YOUR ANSWER" }. Do not introduce information
   beyond the provided sources.
Hungary is a country in [...]
Dorel Moi is a retired [...]
Claudiu Cristian Moldovan is a retired Romanian
\hookrightarrow aerobic gymnast [...]
Claudiu Varlam is a retired Romanian aerobic [...]
Romania is a country located [...]
Sweden , officially the Kingdom of Sweden , is [\ldots]
Russia , or the Russian Federation , is a [\ldots]
France , officially the French Republic
                                             is [...]
South Korea , officially the Republic of Korea is
\hookrightarrow [...]
Germany , constitutionally the Federal Republic of
\hookrightarrow Germany , is [...]
Japan is an island country located in [...]
Italy , officially the Italian Republic , is [...]
```

Expected output:

```
{"answer": "French Republic"}
```

Figure 14: Example for Hybrid QA task (HybridQA).

Table:

Chen Wu-hsiung Minister of the Council of coulture of the Executive Yuan In office 20 May 2008 - 6 February 2012 Preceded by Su Chia-chyuan Succeeded by Chen Bao-ji Deputy Minister of the Council of Agriculture of the Executive Yuan In office 1999-2002 Minister Chen Hsi-huano Fan Chen-tzung Personal details Born 11 March 1944 (age 81) Taihoku, Taiwan, Empir of Japan Republic of China Nationality Political party 🚺 Kuomintang National Chung Hsing University University of Illinois at Alma mater Urbana-Champaign

Instruction:

```
The target cell value in the 1st row and the 1st

→ column is {"row_id":"1", "column_id":"1",

→ "cell_value":"Chen wu-hsiung"}.
```

Figure 15: Example for Table Cell Extraction.

Table: Busiest international routes at Manzanillo International Airport (2013)^[1]

Rank	City	Passengers	Ranking	Airline
1	United States, Los Angeles	14,749	_	Alaska Airlines
2	United States, Houston	5,465	_	United Express
3	■ Canada, Calgary	3,761	_	Air Transat, WestJet
4	I ♦ I Canada, Saskatoon	2,282	4	
5	Canada, Vancouver	2,103	_	Air Transat
6	United States, Phoenix	1,829	1	US Airways
7	Canada, Toronto	1,202	▼ 1	Air Transat, CanJet
8	I ◆ I Canada, Edmonton	110		
9	United States, Oakland	107		

Instruction:

Answer this: How many more passengers flew to los \hookrightarrow angeles than to saskatoon from manzanillo airport \hookrightarrow in 2013? Consult the table and answer. Return your \hookrightarrow answer as JSON: {"answer": "YOUR ANSWER"}. Avoid \hookrightarrow including information not present in the table.

```
{"answer": "12,467"}
```

Figure 16: Example for Table QA task (WikiTableQuestions).

Nationality	Player	Ranking*	Seeding
▼ ISR	Dudi Sela	56	1
TPE	Lu Yen-hsun	62	2
■■ USA	Bobby Reynolds	93	3
GER	Michael Berrer	111	4
BRA	Thiago Alves	120	5
GER	Benjamin Becker	126	6
■ FRA	Nicolas Mahut	137	7
RUS	Michail Elgin	138	8

Instruction:

```
For the table found in '2009 israel open'

- "seeds" section on Wikipedia,

identify the correct column type

labels for the highlighted column

section is given the following type options:

tv.tv_personality, time.event,

american_football.football_team,

[...]. Provide only the chosen

type(s), separated by commas if

multiple, within the list in this

JSON: {"answer": ["ANSWER"]}.
```

```
{"answer": ["location.country",

→ "location.location"]}
```

Figure 17: Example for Colum Type Annotation task based on highlighted table columns.