TIU-Bench: A Benchmark for Evaluating Large Multimodal Models on Text-rich Image Understanding

Kun Zhang, Liqiang Niu , Zhen Cao, Fandong Meng*, Jie Zhou
[1]Pattern Recognition Center, WeChat AI, Tencent Inc
{peterkzhang, zhenzcao}@tencent.com

Abstract

Text-rich images are ubiquitous in real-world applications, serving as a critical medium for conveying complex information and facilitating accessibility. Despite recent advances driven by Multimodal Large Language Models (MLLMs), existing benchmarks suffer from limited scale, fragmented scenarios, and evaluation protocols that fail to fully capture holistic image understanding. To address these gaps, we present **TIU-Bench**, a large-scale, multilingual benchmark comprising over 100,000 full-image annotations and 22,000 rigorously validated question-answer (QA) pairs that span 18 subtasks across diverse real-world scenarios. **TIU-Bench** introduces a novel full-image structured output format that jointly models geometric, textual, and relational information, enabling fine-grained evaluation of perception and reasoning capabilities. Furthermore, we propose a two-stage understanding framework named **T2TIU**, which first generates a structured representation of the entire image and subsequently conducts reasoning on this representation in order to address complex visual-textual queries. Extensive experiments on 10 state-of-the-art generative models highlight the challenges and opportunities in advancing text-rich image understanding. Our benchmark and framework provide a comprehensive platform for developing and evaluating next-generation multimodal AI systems.

1 Introduction

Text-rich images play a pivotal role in real-world scenarios by efficiently conveying complex information and improving accessibility (Biten et al., 2019). Accurate interpretation of such images is essential for automating information extraction, advancing AI systems, and optimizing user interactions. To formalize this research domain, we define **Text-rich Image Understanding (TIU)** as consisting of two core capabilities: *perception* and *understanding*. The perception dimension encompasses visual recognition tasks such as text detection (Liao et al., 2022), text recognition (Guan et al., 2025), formula recognition (Truong et al., 2024; Guan et al., 2024), and document layout analysis (Yupan et al., 2022). In contrast, the understanding dimension involves semantic reasoning for downstream applications such as key information extraction and document-based visual question answering (e.g., DocVQA (Mathew et al., 2021), ChartQA (Masry et al., 2022), and TextVQA (Singh et al., 2019)).

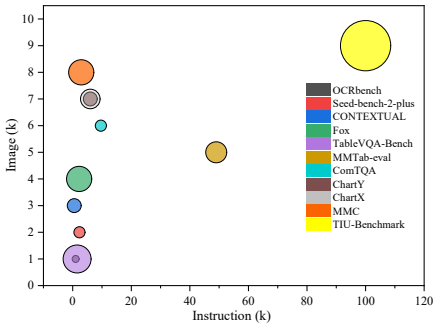

Figure 1: Comparison of the number of images and instructions between our dataset and existing datasets.

Recently, MLLMs have been proposed, which integrate large language models (LLMs) with visual encoders to jointly process visual tokens and linguistic elements through unified attention mechanisms, enabling end-to-end sequence modeling. Within the **TIU** domain, MLLMs have demonstrated impressive results in both perception and understanding. Nevertheless, despite recent advances, there are still two key challenges in current TIU research paradigms.

Dataset Limitations. TIU tasks currently face challenges related to data diversity, scale, and quality. Existing datasets such as DocVQA and ChartQA focus on isolated scenarios (e.g., documents, tables, or charts), and their fragmented objectives and scenario-specific designs hinder the

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24286–24295
November 4-9, 2025

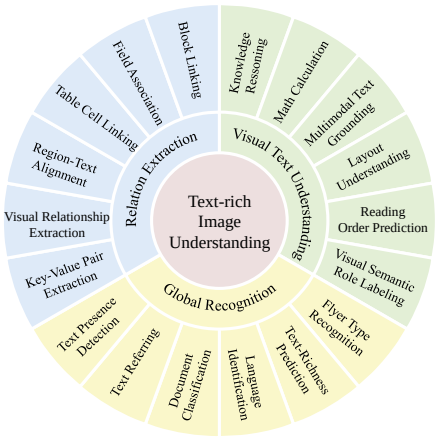

Figure 2: Overview of sub-tasks in TIU-Bench datasets.

comprehensive evaluation of perceptual and interpretative capabilities. Recent benchmarks such as OCRbench and its v2 version assess line-level text recognition in MLLMs but overlook holistic image understanding challenges and suffer from limited language coverage and limited scale. Figure 1 presents a quantitative comparison, illustrating the significantly larger scale and greater diversity of our benchmark compared to existing ones.

Task Setting and Evaluation Constraints. Current evaluation protocols predominantly adopt visual question answering (VQA) formats, which assess question-answer pairs via textual metrics such as edit distance and recall. This approach fails to adequately capture full-image perception due to two key issues: (1) 62% of queries target isolated text regions within complex images, allowing models to circumvent the need for global understanding; (2) 21% of QA pairs reference popular culture entities (e.g., movies, celebrities), which can be solved using textual priors without requiring visual grounding. Our preliminary experiments reveal that leading MLLMs fail to utilize 9.2% of local OCR content during global processing, exposing critical gaps in existing evaluation methodologies.

To address these challenges, we introduce **TIU-Bench**, a comprehensive and rigorously designed benchmark for evaluating language-multimodal models (LMMs) across four literacy dimensions: multilingual text reading (covering 12 languages), text recognition under challenging conditions, layout-aware document parsing, and key information extraction. Comprising 100,213 full-image annotations and 22,000 QA pairs, **TIU-Bench** systematically evaluates three core capabilities (see Figure 2): (1) text recognition (8 sub-tasks including distorted text detection), (2) relation extraction (9 sub-tasks spanning spatial and semantic relationships), and (3) visual-textual reasoning (6 subtasks requiring cross-modal alignment). Each QA pair undergoes rigorous human and large language model cross-validation to ensure annotation quality. Notably, TIU-Bench introduces region-structure-aware reasoning challenges that test MLLMs' ability to analyze hierarchical relationships between image regions (e.g., diagram components in technical manuals), effectively simulating real-world scenarios demanding complex visual-textual understanding.

To enable a fine-grained assessment of holistic image perception, our benchmark adopts a full-image structured output format that simultaneously captures the following : (1) spatial locations of all text segments via bounding box coordinates, (2) recognition results including confidence scores, and (3) inter-paragraph relationships modeled as spatial-semantic graphs. This structured representation underpins our multi-level metrics. Beyond the benchmark, we propose a novel two-stage understanding framework tailored for text-rich images, named **T2TIU**. In the first stage, the model generates a structured representation of the entire image. In the second stage, it reasons over both the complex query and the structured output to derive the final answer. This flexible and extensible approach enables MLLMs to comprehensively capture both holistic scene context and structural relationships among elements, thereby enhancing reasoning capabilities for diverse and complex queries.

In summary, our contributions are fourfold:

- We introduce **TIU-Bench**, a large-scale, multilingual benchmark that evaluates MLLMs' perceptual and comprehension abilities across 18 tasks, 12 languages, and 22 diverse scenarios.
- We propose a robust and comprehensive evaluation suite that measures multimodal responses across multiple performance dimensions, enabling rigorous and fine-grained assessment.
- We propose a two-stage understanding framework, named **T2TIU**, that leverages both global image context and local structural relations to enhance MLLMs' understanding of complex visual inputs.
- We perform an extensive evaluation of 10 state-of-the-art generative models on TIU-

Bench, providing valuable insights into the strengths and limitations of current multimodal understanding approaches.

2 Task Formulation

We formally define the task of **Text-Rich Image Understanding (TIU)**, which aims to analyze images containing rich textual information comprehensively. This task is composed of two interrelated sub-tasks: *Full-Image Parsing* and *Visual Question Answering (VQA)*.

Full-Image Parsing. Given an input image $\mathcal{I}$ that contains dense textual content, the objective of full-image parsing is to extract a structured representation $\mathcal{S}$ of the image. Specifically, $\mathcal{S}$ includes a set of text paragraphs $\{p_1, p_2, \ldots, p_m\}$, each associated with spatial coordinates and orientation angles. Formally, the output can be expressed as:

$$\mathcal{S} = \{(p_i, c_i, \theta_i) \mid i = 1, 2, \ldots, m\},$$

where p_i denotes the i-th textual paragraph detected in the image, c_i represents its bounding box coordinates, and θ_i indicates the text orientation angle relative to a reference axis.

Visual Question Answering (VQA). The VQA sub-task takes as input both the image $\mathcal{I}$ and a natural language question q related to the textual content within the image. The goal is to generate a precise and contextually relevant answer a:

$$a = \mathcal{F}(\mathcal{I}, q),$$

where $\mathcal{F}$ denotes the VQA model that leverages the multimodal information embedded in the image and the semantics of the question to produce the answer.

Together, these two sub-tasks enable a holistic understanding of text-rich images by not only parsing and structuring the textual content but also supporting interactive question answering grounded in the image context.

3 Dataset Construction

In this section, we present the construction details of TIU-Bench.

The process of creating datasets in different domains can be divided into three stages: (1) Data Selection and Preprocessing, (2) QA Generation and Refinement, (3) Data Quality Check. The overview of this process is shown in Figure 3.

3.1 Data Selection and Preprocessing

We undertake three steps during this stage: (1) Data Collection, (2) Data Filtering and Annotation, and (3) Data Refinement.

3.1.1 Data Collection

The goal of this step is to collect TIU task-oriented multimodal data.

Web Data. This category includes Wikipedia pages and articles, derived from Wit dataset (Srinivasan et al., 2021), WikiWeb2M dataset (Burns et al., 2023), and WebQA dataset (Chang and Bisk, 2022). In addition, we also collected freely available images from numerous news and image websites to further enrich our dataset.

Multi-Scene Data. We collect a diverse set of multi-scene datasets, primarily sourced from well-established academic benchmarks, including documents (FUNSD (Jaume et al., 2019) and IAM (Marti and Bunke, 2002)), multi-orientation text, and artistic text. In addition, we incorporate several newly collected datasets covering street scenes (Scene-zh), web scenes, HierAgent, and LAION-OCR. To ensure comprehensive scenario coverage, we further augment these with proprietary private data. Altogether, our dataset spans 22 representative scenarios, including schematic diagrams, scientific papers, text image patches, filled tables, charts, receipts, question contexts, mathematical formulas, product labels, phone screenshots, indoor scenes, and more.

Multilingual Data. We collect multilingual data from social media platforms across various languages, as well as from publicly available datasets. For each language, the dataset includes 10 real-shot document images and 1,400 natural scene images. The document images are entirely newly collected, featuring multi-orientation and real-world captures. The natural scene images are partly re-annotated from the MTVQA dataset (Tang et al., 2024) and partly sourced from newly collected data, with a particular focus on Russian, Spanish, and Portuguese.

3.1.2 Data Filtering and Annotation

The goal of this step is to clean the data and obtain high-quality annotations, ensuring that images correctly correspond to their structured outputs.

We first filter out data entries that lack image information or contain excessively large images. Next, we apply two stages of deduplication: (1)

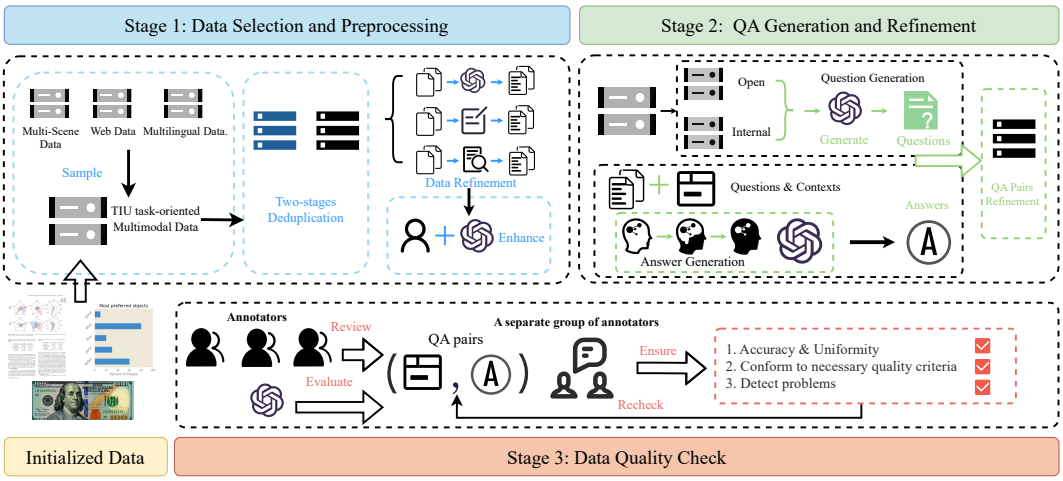

Figure 3: The overview of TIU-Benchmark dataset construction.

MinHash-based deduplication (Shrivastava and Li, 2014) to remove data with visually similar images, and (2) string and semantic similarity measures (Xiao et al., 2024) to further eliminate duplicate entries.

Subsequently, we construct prompts to have GPT-4o and Qwen-2.5-VL-72B generate full-image analyses for each image. For every image, we perform cross-validation of the outputs from both models. Images and their corresponding full-image analysis results that demonstrate consistent agreement after multiple rounds of validation are retained.

3.1.3 Data Refinement

The final step in this stage is to refine the image-text interleaved contexts, along with the questions and answers present in the original data. We employ both GPT-4o and human verification to ensure quality.

3.2 QA Generation and Refinement

This stage is consisted of three steps: (1) Question Generation, (2) Answer Generation, (3) QA Pair Refinement.

3.2.1 Question Generation

We leverage GPT-4o to generate high-quality, context-specific questions tailored to the provided images. We follow two key criteria: (1) questions are constructed based on the corresponding contexts, and (2) questions should be natural and practical, with the potential to be effectively answered through the integration of images.

3.2.2 Answer Generation

For each sample, given the generated or original questions Q and corresponding image I, we generate image-text pairs using GPT-4o following a CoT reasoning strategy (Wei et al., 2022): (1)**Question Validity Assessment**: we first ensure that valid questions $ValQ$ can be directly answered by the images I, and remove invalid questions $InvQ$. (2) **Evidence Extraction**: we then extract evidence E from the images I to support the answer A for subsequent answer construction. (3) **Answer Construction**,finally, we construct a highly reliable, accurate, and coherent answer based on the valid questions $ValQ$ and the extracted evidence E.

3.2.3 QA Pair Refinement

We further refine the QA pairs formulated in the previous steps following the optimization approach outlined by Zhu et al. (2024); Yu et al. (2025). We leverage GPT-4o to extract supporting evidence from contexts and verify its alignment with keywords in the answer.

3.3 Data Quality Check

To ensure the high quality and reliability of TIU-Bench, we further verify the consistency and correctness of image-structural output pairs and QA pairs. Multiple rounds of sampling are performed on the datasets, and annotators are engaged in a structured data quality check process. This process consists of three steps: (1) **Initial Review:** A group of annotators reviews sampled image-structural output pairs and QA pairs, assessing their correctness and consistency. They identify any problematic entries for further revision.(2) **Issue Correction:** A separate group of annotators addresses the iden-

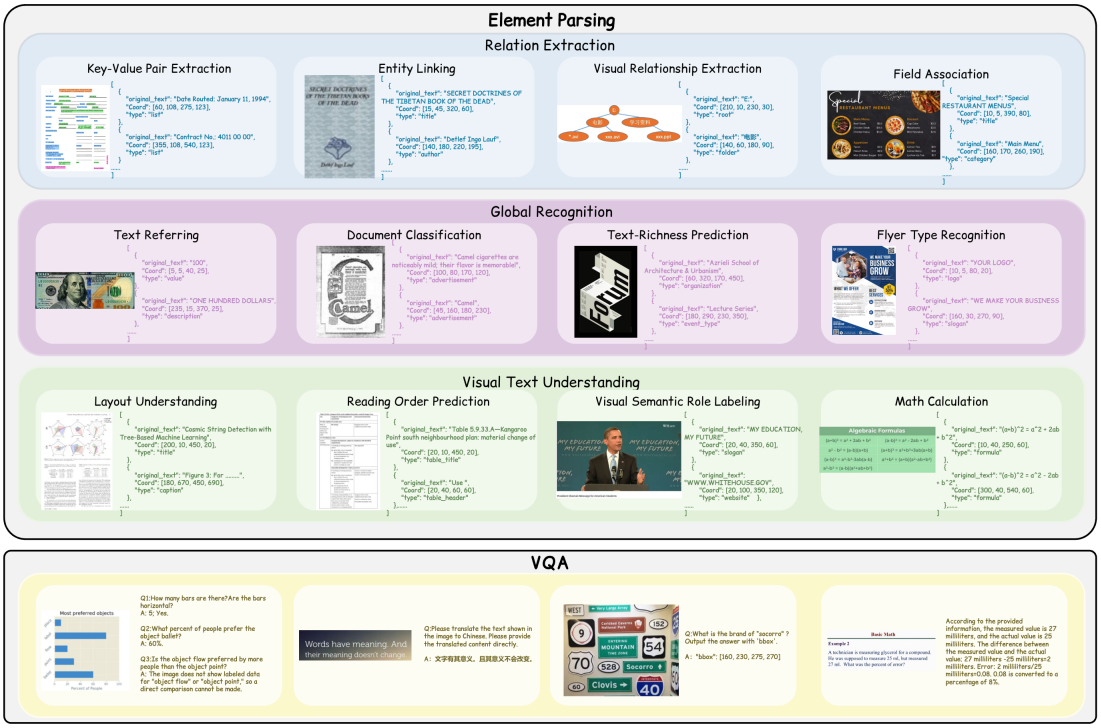

Figure 4: Sample visualizations for sub-tasks in TIU Benchmark Datasets.

tified issues, including errors in structural outputs, question formulations, and answer correctness. Expert annotators conduct a meticulous review and correction across the datasets.(3) **Final Verification:** A team of reviewers performs a comprehensive recheck to ensure overall dataset accuracy. This step validates that all corrections have been properly implemented and that the datasets meet the required quality standards.

3.4 Data Statistics

Based on the aforementioned construction pipeline, the TIU benchmark dataset comprises 100,213 images along with their corresponding full-image parsing annotations. Additionally, it includes a carefully curated set of 22,000 question-answer pairs, which cover 18 distinct sub-tasks.

4 Framework

In this section, we present our framework, which consists of two stages: (1) Holistic Image Parsing, and (2) the generation of answers based on the parsing results, utilizing a foundational generative model.

4.1 Holistic Image Parsing

The first stage employs a Multimodal Large Language Model (MLLM) to perform a comprehensive parsing of the input image $\mathcal{I}$. The goal is to extract a structured representation $\mathcal{S}$ that encodes both the semantic and spatial information of the textual content within the image.

Formally, given an input image $\mathcal{I}$, the MLLM produces a set of N detected text segments (paragraphs):

$$\mathcal{S} = \{(p_i, b_i, \theta_i) \mid i = 1, 2, \ldots, N\},$$

where each element consists of:

- p_i: the recognized text content of the i-th paragraph,
- $b_i = (x_i, y_i, w_i, h_i)$: the bounding box coordinates representing the position and size of the paragraph in the image, with (x_i, y_i) denoting the top-left corner, and w_i, h_i the width and height respectively,
- θ_i: the orientation angle of the paragraph relative to the image coordinate system, capturing possible rotations.

The parsing process can be viewed as a mapping function:

$$MLLM : \mathcal{I} \mapsto \mathcal{S},$$

which jointly performs text detection, recognition, and geometric estimation.

To further capture the spatial relationships among paragraphs, we model the inter-paragraph relations as a spatial semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the vertex set $\mathcal{V} = \{v_i\}_{i=1}^{N}$ corresponds to the detected paragraphs, and the edge set $\mathcal{E}$ encodes spatial or semantic relations such as adjacency, reading order, or hierarchical structure. Formally, each edge $e_{ij} \in \mathcal{E}$ can be represented as:

$$e_{ij} = f_{rel}(b_i, b_j, p_i, p_j),$$

where $f_{rel}(\cdot)$ is a learned or heuristic function that quantifies the relationship between paragraph i and paragraph j based on their spatial layout and textual content.

Thus, the holistic parsing output can be compactly represented as:

$$\mathcal{S} = \{(p_i, b_i, \theta_i, c_i)\}_{i=1}^{N}, \quad \mathcal{G} = (\mathcal{V}, \mathcal{E}),$$

providing a rich, structured, and interpretable representation of the image content that serves as the foundation for subsequent reasoning and answer generation.

4.2 Answer Generation

In the second stage, the framework generates answers conditioned on the input image, the structured parsing results from the first stage, and the posed question. By integrating the visual information with the structured semantic layout, the model is able to perform more accurate and context-aware reasoning to produce the final answer.

Formally, given an input image $\mathcal{I}$ and a question q, the holistic parsing stage produces a structured output $\mathcal{S} = MLLM(\mathcal{I})$, which includes paragraph texts, bounding boxes, and angles. Subsequently, the answer generation stage utilizes the tuple $(\mathcal{I}, \mathcal{S}, q)$ to generate the answer $\mathcal{A}$, i.e.,

$$\mathcal{S} = MLLM(\mathcal{I}), \quad \mathcal{A} = Gen(\mathcal{I}, \mathcal{S}, q),$$

where $Gen(\cdot)$ denotes the answer generation model.

This two-stage design effectively decouples the complex task of rich text image understanding into interpretable parsing and reasoning components, enabling improved performance and interpretability on the TIU benchmark.

5 Evaluation Metrics

To achieve a fine-grained evaluation of holistic image understanding, our benchmark adopts a comprehensive structured output format that simultaneously captures: (1) the geometric locations of all text segments via bounding box coordinates, (2) recognition results including confidence scores, and (3) inter-paragraph relationships modeled as a spatial semantic graph. This structured representation supports our multi-level evaluation protocol:

Global Evaluation: Missing text detection to identify unrecognized paragraphs. In this part, we generate a structured output for the entire image and then compare it against the globally annotated OCR in the dataset to determine the proportion of missed paragraphs. This metric is referred to as *OCR-complete*.

Local Evaluation: (a) Visual alignment is measured by the intersection over union (IoU) between predicted and ground truth bounding boxes to assess coordinate accuracy. We compute Recall, Precision, and F1 scores for both paragraph coordinates and recognized text, denoted as *Para-Recall*, *Para-Precision*, and *Para-F1*, respectively; (b) Text accuracy is quantified at the character level using the Levenshtein distance to measure consistency-denoted as *Avg-Edit-Distence*.

QA Evaluation: We evaluate the QA task using Recall and Precision metrics, referred to as *QA-Recall* and *QA-Precision*, respectively.

6 Experiments

6.1 Experimental Baselines and Settings

Baselines We evaluate 2 closed-source models and 8 open-source models. Specifically, we select 2 popular closed-source multimodal large models: GPT-4o (Achiam et al., 2023), GPT-4o-mini (Achiam et al., 2023). For the open-source models, we choose 9 multimodal large models: Qwen2.5-VL-3B-Instruct (Wang et al., 2024), Qwen2.5-VL-7B-Instruct, Qwen2.5-VL-32B-Instruct, Qwen2.5-VL-72B-Instruct (Wang et al., 2024), LLaVA-Next-8B (Chen et al., 2024b), LLaVA-OV-7B(Li et al., 2024a), InternVL2-8B (Chen et al., 2024c) and InternVL2-26B (Chen et al., 2024c).

Experiment Details For open-source large models with fewer than 40B parameters, we adopt supervised fine-tuning. Specifically, we split the Full-image Parsing and Visual Question Answering datasets into training, validation, and test sets with an 8:1:1 ratio. The models are trained on the training set, the best model parameters are selected based on validation performance, and a final evaluation is conducted on the test set.

For larger models such as Qwen-2.5-VL-72B, Closed-Source GPT-4o, and GPT-4o-mini, we em-

Method	Global-Eval	Local-Eval				QA-Eval	
	OCR-Complete	Para-Recall	Para-Precision	Para-F1	Avg-Edit-Distence	QA-Recall	QA-Precision
Open-source LMMs							
Qwen2.5-VL-3B (Bai et al., 2023)	83.2	73.4	76.8	75.2	4.42	63.2	67.3
Qwen2.5-VL-7B (Bai et al., 2023)	87.4	77.2	79.4	78.6	3.92	67.5	69.2
Qwen2.5-VL-32B (Wang et al., 2024)	91.2	87.2	90.3	88.9	1.34	71.6	73.4
Qwen2.5-VL-72B (Wang et al., 2024)	93.4	85.2	94.3	92.5	1.02	76.2	80.3
LLaVA-Next-8B (Liu et al., 2024b)	85.6	75.8	77.2	76.3	4.03	66.3	67.2
LLaVA-OV-7B (Li et al., 2024a)	83.8	76.8	78.2	77.8	3.86	65.8	67.2
InternVL2.5-8B (Chen et al., 2024b)	79.6	77.4	79.2	78.5	3.63	67.8	70.3
InternVL2.5-26B (Chen et al., 2024b)	88.9	84.2	86.6	85.9	1.55	72.1	74.5
Dolphin-322M	87.2	76.6	78.6	79.2	3.99	72.1	74.5
Pixtral-12B	88.5	84.8	89.6	87.2	1.52	72.1	74.5
Closed-source LMMs							
GPT-4o (Achiam et al., 2023)	95.3	92.6	95.1	94.2	0.82	80.2	82.6
GPT-4o-mini (Achiam et al., 2023)	92.0	88.6	90.2	89.6	1.10	73.5	76.8
LMM with Our Framework							
Qwen2.5-VL-3B+T2TIU	83.2	73.4	76.8	75.2	4.42	65.8	70.2
Qwen2.5-VL-7B+T2TIU	87.4	77.2	79.4	78.6	3.92	69.6	72.1
LLaVA-Next-8B+T2TIU	85.6	75.8	77.2	76.3	4.03	68.9	70.1
LLaVA-OV-7B+T2TIU	83.8	76.8	78.2	77.8	3.86	65.8	67.2
InternVL2.5-8B+T2TIU	79.6	77.4	79.2	78.5	3.63	70.4	73.1
GPT-4o-mini+T2TIU	92.0	88.6	90.2	89.6	1.10	75.8	78.2

Table 1: **Evaluation of existing methods on Full Image Parsing and Visual Question Answering**. The notations apply to all subsequent figures.

ploy a few-shot learning approach, providing the model with a minimal number of samples for training.

6.2 Experiment Results

6.2.1 Full-Image Parsing Performance

As shown in Table 1, we observed that all models tend to miss certain OCR instances, a phenomenon that was not reflected in previous datasets and task settings. This highlights the advanced nature of our dataset and task design in the context of the TIU-Bench.

As the model size increases within the same series, the OCR omission rate gradually decreases, while local evaluation metrics—namely paragraph-level recall, precision, and F1 score—show steady improvement. From the full-image parsing results, we observe that even the currently popular large multimodal models exhibit limitations in their full-image parsing capabilities. For models with fewer than 10 billion parameters, the paragraph-level metrics remain relatively low, which often constrains their performance on downstream reasoning and other related tasks. When a model struggles to accurately interpret the information within an image, subsequent tasks tend to suffer from error accumulation in a snowball effect. Correspondingly, we also observe that as the model's global parsing ability improves, its reasoning capabilities progressively strengthen as well.

6.2.2 Visual Question Answering Performance

In this section, we evaluate Visual Question Answering Performance with results presented for our datasets.

As shown in Table 1, advanced models such as GPT-4o and GPT-4o-mini consistently outperform smaller open-source models ($\sim$ 7B parameters) across all domains and methods. These smaller models exhibit subpar performance across different methods and dataset domains, even when utilizing rule-based generation techniques. In contrast, larger open-source models ($\sim$ 70B parameters) significantly reduce the performance gap with closed-source models.

On more challenging datasets, however, the performance gap becomes more pronounced, highlighting the limitations of open-source models in handling complex TIU tasks. Nevertheless, smaller open-source models remain a cost-effective solution for simpler applications with limited computational resources.

6.2.3 The Performance of T2TIU

From Table 1, we can see that our framework improves performance across all models on the TIU-Bench dataset, validating the effectiveness of our approach. Additionally, we observe that the performance gains are more significant for smaller models. This aligns with our previous analysis that larger models inherently possess stronger full-image parsing and reasoning capabilities, while our framework, through a Chain-of-Thought (CoT)-like method, assists models in reasoning based on comprehensive image information.

Furthermore, we conducted experiments on two commonly used datasets in the text and table domains. Our framework consistently enhances performance across all models on these datasets as well, demonstrating its strong generalization ability.

Method	DocVQA	ChartQA
Qwen-2.5VL-3B	89.2	81.0
Qwen-2.5VL-3B+T2TIU	**91.8**	**83.4**
Qwen-2.5VL-7B	91.5	82.6
Qwen-2.5VL-7B+T2TIU	**93.3**	**84.1**

Table 3: The performance of T2TIU on two different datasets.

7 Related Work

To evaluate LMMs, developing a comprehensive benchmark is essential. Previous effort has focused on creating scenario-specific benchmarks to assess LMMs in particular contexts. For example, DocVQA (Mathew et al., 2021) is designed to evaluate the document comprehension abilities of LMMs, while ChartQA (Masry et al., 2022) is tailored to chart interpretation skills. Similarly, Infographics VQA (Mathew et al., 2022) is dedicated to assessing the understanding of infographic images. Additionally, TextVQA (Singh et al., 2019) aims to evaluate text comprehension in real-world scenes. To further investigate the robustness of the model, some methods expand the scope of evaluation scenarios. OCRBench (Liu et al., 2023) introduces a holistic evaluation framework that covers five core OCR tasks. CONTEXTUAL (Wadhawan et al., 2024) is developed with context-sensitive instructions. SEED-Bench-2-Plus (Li et al., 2024b) encompasses a wide spectrum of text-rich images from various sources, including web content, maps, and charts. To provide a more thorough assessment, some benchmarks design multiple evaluation tasks within a specific scenario. TableVQA-Bench (Kim et al., 2024) first focuses on VQA tasks in the table domain. MMTab (Zheng et al., 2024) and ComTQA (Zhao et al., 2024) then extend the task scope, including table detection, structure recognition, and table querying. Moreover, ChartY (Chen et al., 2024a), ChartX (Xia et al., 2024), and MMC (Liu et al., 2024a) evaluate LMMs in chart understanding through tasks such as chart information extraction and reasoning. In this work, we focus on establishing a new benchmark called *TIUBechmarks*, which contains more tasks than previous benchmarks and provides a systematic evaluation framework to reveal the limitations of LMMs in diverse text-rich environments.

8 Conclusion

We present **TIU-Bench**, a large-scale multilingual benchmark addressing key limitations of existing datasets by providing over 100,000 full-image annotations and 22,000 validated QA pairs across 18 diverse subtasks. Our novel structured output format enables fine-grained evaluation of perception and reasoning in text-rich images. Alongside, we propose a two-stage framework that generates structured image representations and performs reasoning to answer complex queries. Experiments on 10 state-of-the-art models reveal significant challenges, underscoring TIU-Bench's value as a comprehensive platform for advancing multimodal AI research.

9 Limitation

While **TIU-Bench** significantly advances the evaluation of text-rich image understanding by providing large-scale, multilingual, and diverse annotations with a novel structured output format, several limitations remain. First, despite the extensive coverage of 18 subtasks, the benchmark may not fully encompass all possible real-world scenarios, especially those involving highly specialized or domain-specific text-image interactions. Second, the two-stage understanding framework, although effective, relies on accurate, structured representation generation as a prerequisite, which may propagate errors and limit end-to-end performance.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. pages 4291–4301.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. Wikiweb2m: A page-level multimodal wikipedia dataset. *arXiv preprint arXiv:2305.05432*.

Yingshan Chang and Yonatan Bisk. 2022. Webqa: A multimodal multihop neurips challenge. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 232–245. PMLR.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. 2024a. Onechart: Purify the chart structural extraction via one auxiliary token. pages 147–155.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. pages 24185–24198.

Tongkun Guan, Chengyu Lin, Wei Shen, and Xiaokang Yang. 2024. Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*, pages 130–147. Springer.

Tongkun Guan, Wei Shen, and Xiaokang Yang. 2025. Ccdplus: Towards accurate character to character distillation for text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. volume 2, pages 1–6. IEEE.

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

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Bohao Li, Yuying Ge, Yi Chen, Yixiao Ge, Ruimao Zhang, and Ying Shan. 2024b. Seed-bench-2-plus: Benchmarking multimodal large language models with text-rich visual comprehension. *arXiv preprint arXiv:2404.16790*.

Minghui Liao, Zhisheng Zou, Zhaoyi Wan, Cong Yao, and Xiang Bai. 2022. Real-time scene text detection with differentiable binarization and adaptive scale fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):919–931.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2024a. MMC: Advancing Multimodal Chart Understanding with Large-scale Instruction Tuning. pages 1287–1310.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Yuliang Liu, Zhang Li, Biao Yang, Chunyuan Li, Xucheng Yin, Cheng-lin Liu, Lianwen Jin, and Xiang Bai. 2023. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

U-V Marti and Horst Bunke. 2002. The iam-database: an english sentence database for offline handwriting recognition. 5:39–46.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. pages 2200–2209.

Anshumali Shrivastava and Ping Li. 2014. In defense of minhash over simhash. In *Artificial Intelligence and Statistics*, pages 886–894. PMLR.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. pages 8317–8326.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, Chunhui Lin, Wanqing Li, Mohamad Fitri Faiz Bin Mahmood, Hao Feng, Zhen Zhao, and 1 others. 2024. Mtvqa: Benchmarking multilingual text-centric visual question answering. *arXiv preprint arXiv:2405.11985*.

Thanh-Nghia Truong, Cuong Tuan Nguyen, Richard Zanibbi, Harold Mouchère, and Masaki Nakagawa. 2024. A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and gnn models. *Pattern Recognition*, 153:110531.

Rohan Wadhawan, Hritik Bansal, Kai-Wei Chang, and Nanyun Peng. 2024. ConTextual: Evaluating Context-Sensitive Text-Rich Visual Reasoning in Large Multimodal Models.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. 2025. Mramg-bench: A beyondtext benchmark for multimodal retrieval-augmented multimodal generation. *arXiv preprint arXiv:2502.04176*.

Huang Yupan, Lv Tengchao, Cui Lei, Lu Yutong, and Wei Furu. 2022. LayoutLMv3: pre-training for document AI with unified text and image masking. *arXiv*, 2204.08387.

Weichao Zhao, Hao Feng, Qi Liu, Jingqun Tang, Shu Wei, Binghong Wu, Lei Liao, Yongjie Ye, Hao Liu, Houqiang Li, and 1 others. 2024. TabPedia: Towards Comprehensive Visual Table Understanding with Concept Synergy. *arXiv preprint arXiv:2406.01326*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal Table Understanding. pages 9102–9124. Association for Computational Linguistics.

Kunlun Zhu, Yifan Luo, Dingling Xu, Ruobing Wang, Shi Yu, Shuo Wang, Yukun Yan, Zhenghao Liu, Xu Han, Zhiyuan Liu, and 1 others. 2024. Rageval: Scenario specific rag evaluation dataset generation framework. *arXiv preprint arXiv:2408.01262*.

A Example Appendix

This is an appendix.