Enhancing Table Recognition with Vision LLMs: A Benchmark and Neighbor-Guided Toolchain Reasoner

Yitong Zhou[1] , **Mingyue Cheng**[1*] , **Qingyang Mao**[1] , **Jiahao Wang**[1] , **Feiyang Xu**[2] , **Xin Li**[1,2]
[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Artificial Intelligence Research Institute, iFLYTEK Co., Ltd
{yitong.zhou, maoqy0503, jiahao.wang}@mail.ustc.edu.cn, {mycheng, leexin}@ustc.edu.cn,
fyxu2@iflytek.com

Abstract

Pre-trained foundation models have recently made significant progress in table-related tasks such as table understanding and reasoning. However, recognizing the structure and content of unstructured tables using Vision Large Language Models (VLLMs) remains under-explored. To bridge this gap, we propose a benchmark based on a hierarchical design philosophy to evaluate the recognition capabilities of VLLMs in training-free scenarios. Through in-depth evaluations, we find that low-quality image input is a significant bottleneck in the recognition process. Drawing inspiration from this, we propose the **N**eighbor-**G**uided **T**oolchain **R**easoner (NGTR) framework, which is characterized by integrating diverse lightweight tools for visual operations aimed at mitigating issues with low-quality images. Specifically, we transfer a tool selection experience from a similar neighbor to the input and design a reflection module to supervise the tool invocation process. Extensive experiments on public datasets demonstrate that our approach significantly enhances the recognition capabilities of the vanilla VLLMs. We believe that the benchmark and framework could provide an alternative solution to table recognition. The code is available at https://github.com/lqzxt/NGTR.

1 Introduction

Tables are ubiquitous for organizing and communicating structured data across diverse domains, ranging from scientific literature and business reports to web pages and financial documents [Ye *et al.*, 2024; Zheng *et al.*, 2021]. They store a wealth of information essential for applications such as knowledge discovery, decision support, and data-driven analytics [Shwartz-Ziv and Armon, 2022; Wang *et al.*, 2024a]. In the context of intelligent table applications, one fundamental yet challenging task is table recognition: converting image-based table representations into structured data formats. Over the years, substantial efforts [Salaheldin Kasem *et al.*, 2024] have been made to address this problem, introducing various approaches to address challenges, such as image segmentation techniques and cell object detection methods.

Recently, the advent of Large Language Models (LLMs) [Chang *et al.*, 2024] and Vision Large Language Models (VLLMs) [Yin *et al.*, 2024] has revolutionized natural language processing and computer vision. For LLMs, their powerful understanding and reasoning capabilities have facilitated numerous tabular data mining tasks, such as table-to-text generation [Guo *et al.*, 2024], table question answering [Wang *et al.*, 2024b; Mao *et al.*, 2024], and table semantic understanding [Deng *et al.*, 2022; Cheng *et al.*, 2025a]. Meanwhile, several VLLM-based methods have emerged to bypass traditional OCR pipelines for visual table analysis and understanding [Hu *et al.*, 2024].

Despite these advancements, our investigation reveals a noticeable gap: the application of VLLMs to table recognition remains underexplored. This task serves as a foundational building block for table-related applications. Some existing work [Luo *et al.*, 2024; Zhang *et al.*, 2024] has focused on pre-training or fine-tuning VLLMs to accomplish this task. However, fine-tuning VLLMs for specific tasks is often computationally expensive and risks catastrophic forgetting of general capabilities. To address this, we explore a generative approach that does not require additional fine-tuning, specifically leveraging a training-free paradigm using pre-trained VLLMs for table recognition. Recognizing the absence of dedicated benchmarks in this domain, we propose an evaluation benchmark based on a hierarchical design philosophy [Sui *et al.*, 2024; Cheng *et al.*, 2025b; Liu *et al.*, 2024b], comprising recognition tasks for table recognition. Through extensive evaluations, we identify a critical bottleneck: low-quality input images significantly hinder the table recognition capabilities of the evaluated VLLMs.

To overcome this limitation, we propose the Neighbor-Guided Toolchain Reasoner (NGTR) framework for effective table recognition. One of the key features of the framework is its integration of lightweight models and the strategy of retrieval-augmented generation to improve image quality and guide structured data recognition. Specifically, we propose a preprocessing toolkit with various lightweight models to enhance input image quality. For each input instance, we retrieve a similar neighbor from the training data and use the experience gained from that neighbor to guide the generation of tool invocation plans. Furthermore, we incorporate

*Corresponding author

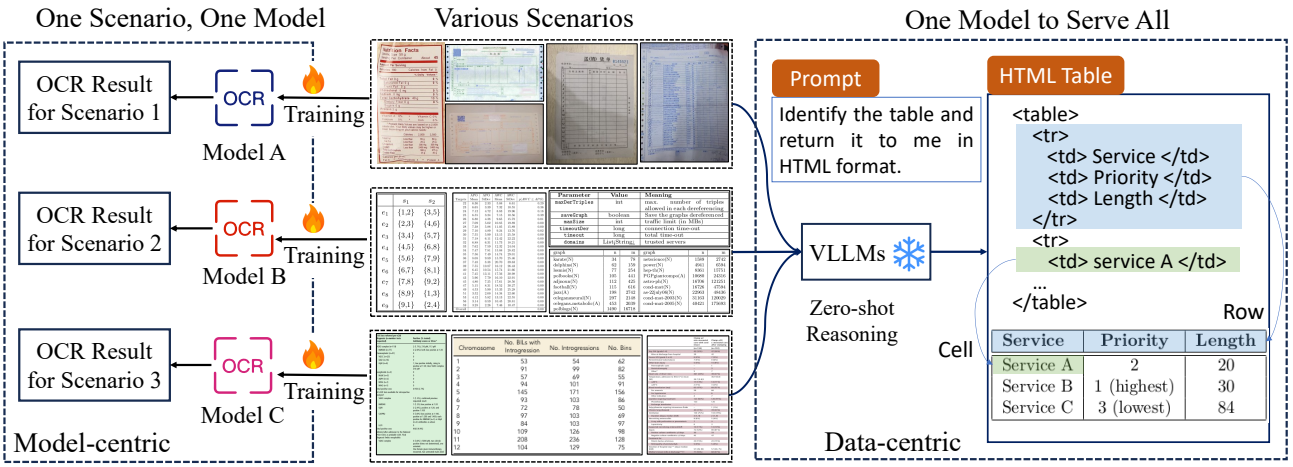

Figure 1: Comparison of modeling paradigms: domain-specific lightweight models vs. universal pre-trained VLLMs.

a reflection-driven tool selection module at each step to iteratively refine the table recognition output. This enables VLLMs to produce more accurate structured data.

To validate the effectiveness of the proposed NGTR framework, we conduct extensive experiments on multiple public table recognition datasets. The key observations are as follows: (1) Our NGTR framework significantly enhances the table recognition performance of naive VLLM-based approaches; (2) While VLLMs achieve competitive accuracy on specific datasets compared to traditional models, a noticeable performance gap remains in favor of traditional models. Nonetheless, we have preliminarily revealed the performance boundaries of VLLMs in several representative table recognition datasets. As is shown in Figure 1, the VLLM-based table recognition approach demonstrates the capability for universal modeling. This method facilitates a paradigm shift in design objectives from a model-centric to a data-centric focus, presenting significant potential for further exploration. We hope this work will inspire more research efforts in the future. In summary, the contributions of this paper are as follows:

- We conduct a systematic investigation into VLLM-based table recognition by introducing a hierarchical benchmark for evaluating their recognition capabilities.
- We propose the NGTR framework to address critical bottlenecks in table recognition, such as low-quality input images.
- We conduct extensive experiments to report the promising performance and potential of VLLMs for table recognition, along with interesting observations that highlight areas for future research.

2 Related Work

Table Recognition. Earlier table recognition (TR) methods predominantly rely on heuristic rules [Kieninger and Dengel, 1999; Shigarov *et al.*, 2016], these approaches rely heavily on handcrafted features or implicit rules. In the era of deep learning, numerous studies have made impressive progress in handling more intricate and heterogeneous table structures. *Top-down methods* [Schreiber *et al.*, 2017; Siddiqui *et al.*, 2019; Ma *et al.*, 2023; Qin *et al.*, 2024] predict table borders to infer the structure information. *Bottom-up methods* [Zheng *et al.*, 2021; Qiao *et al.*, 2021; Xing *et al.*, 2023] first identify table cells with object detection models [Ren *et al.*, 2016; Carion *et al.*, 2020], and then predict the cell relations to organize the row-column structures to form the overall tables. These methods follow an explicit two-stage learning paradigm with relatively strong transferability and explainability, yet the risks of ambiguous contents and boundless structures may lead to unstable and incorrect prediction results. Recently, *sequence-based methods* [Zhong *et al.*, 2020; Nassar *et al.*, 2022; Huang *et al.*, 2023] have been widely explored to directly generate markup sequences that define structures with specific decoders. Although these approaches require a massive of training data and computing resources, they have demonstrated substantial potential to unify visual-text parsing tasks [Wan *et al.*, 2024].

Large Language Models. In recent years, LLMs have demonstrated exceptional performance in tasks such as multi-task learning [Chen *et al.*, 2024], zero-shot learning [Kojima *et al.*, 2022], and text generation [Li *et al.*, 2024]. LLMs have not only broken through the limitations of traditional technologies in processing natural language text, but have also shown capabilities in reasoning [Wei *et al.*, 2022] and planning [Guan *et al.*, 2023; Gou *et al.*, 2023]. Meanwhile, VLLMs combine visual and language understanding capabilities, enabling LLMs to process visual information. For multimodal understanding scenarios (*e.g.*, scene text recognition [Wang *et al.*, 2011], visual question answering [Antol *et al.*, 2015]), VLLMs have been widely validated as effective [Guo *et al.*, 2023; Ye *et al.*, 2023; Liu *et al.*, 2024a]. With continuous progress in text-rich scenarios, some studies [Zheng *et al.*, 2024; Zhao *et al.*, 2024; Chen *et al.*, 2023] have also focused on enabling VLLMs to handle multimodal table understanding tasks.

Despite their promising success in various domains, VLLMs applied to TR remain under-evaluated and under-explored. Our study presents a comprehensive benchmark

Granularity	Recognition Task	Description
Table-level	Visual Table Size Detection	Get the number of rows and columns.
Row-level	Row Index-based Data Recognition	Get the content list of a specific row.
Column-level	Column Index-based Data Recognition	Get the content list of a specific column.
Cell-level	Merged Cell Detection	Get contents of all merged cells.
	Content-based Cell recognition	Get the location of specific cell content.
	Index-based Cell Recognition	Get the cell content of specific location.

Table 1: Descriptions of the proposed hierarchical recognition tasks.

for VLLM-based TR evaluation. Subsequently, we propose a novel framework to address the bottleneck of VLLMs, thereby enhancing their capabilities in TR.

3 Preliminary and Proposed Benchmark

3.1 Problem Definition

We employ the generation paradigm of VLLMs to address the table recognition (TR) task, which is formulated as a format mapping problem from images to sequences. Formally, given a TR dataset $\mathcal{D} = \{(I^i, H^i)\}_{i=1}^n$ with n samples, we predict the corresponding structured form H^i for each table image I^i. Specifically, we provide the image table I^i along with a prompt P as input to the VLLMs, which generates the structured data form $\hat{H}^i = \text{VLLM}(P, I^i)$.

3.2 Benchmark Evaluation Setup

This section proposes an evaluation benchmark for TR based on VLLMs, outlining the hierarchical recognition tasks to assess their performance and the evaluation setup.

Recognition Task Design

We design several hierarchical recognition tasks to conduct a more in-depth assessment of VLLMs' table recognition capability, including the cell-level, row-level, column-level, and global table-level. Table 1 presents the details of the hierarchical recognition tasks.

Cell-level. We evaluate the cell-level recognition capability of VLLMs within three specific recognition tasks. *Merged cell detection* task aims to recognize cells that span multiple rows or columns in the table. *Content/index-based cell recognition* tasks evaluate the structural and content recognition capabilities of VLLMs, which are crucial for assessing whether VLLMs could perform well in fine-grained table recognition.

Row/Column-level. We evaluate the row/column-level recognition capability of VLLMs within two specific recognition tasks. *Row/column index-based data recognition* tasks are designed to assess whether VLLMs can accurately identify row/column elements.

Table-level. We evaluate the overall table-level recognition capability of VLLMs. The *visual table size detection* recognition task is designed to evaluate whether VLLMs can comprehend the global structural information and accurately determine the number of rows and columns in a table.

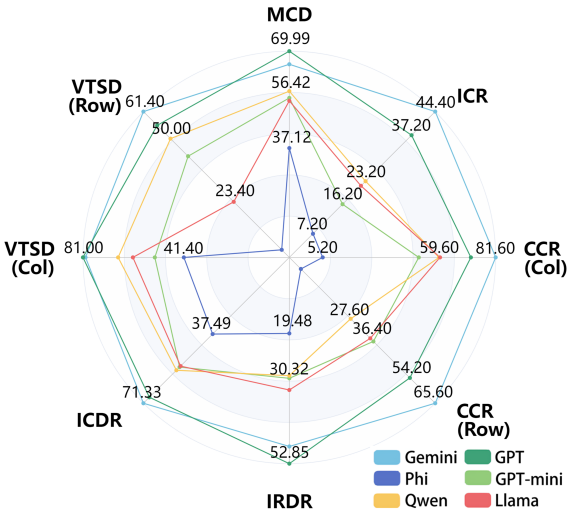

Figure 2: Experimental results of VLLMs for the proposed hierarchical tasks. The tasks evaluated include the following: merged cell detection (MCD), content/index-based cell recognition (CCR, ICR), index-based row/column data recognition (IRDR, ICDR), and visual table size detection (VTSD).

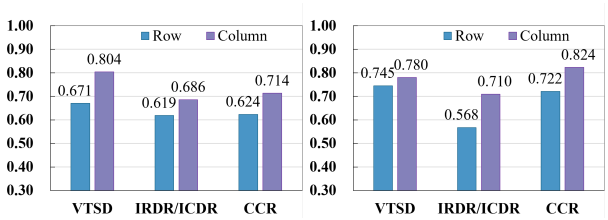

Figure 3: Row-Column Sensitivity Analysis of VLLMs on Hierarchical Tasks with Gemini (Left) and GPT (Right).

Baselines

In this study, we evaluate the performance of six VLLMs. For open-source VLLMs, we select Phi-3.5 (Phi) and Llama-3.2-90B (Llama) for evaluation. For closed-source VLLMs, we evaluat GPT-4o-mini (GPT-mini), Qwen-VL-Max (Qwen), GPT-4o (GPT) and Gemini-1.5-Pro (Gemini).

3.3 Benchmark Evaluation Results

Evaluation results of hierarchical recognition tasks are presented in Figure 2. Among all the VLLMs we selected, GPT and Gemini demonstrate the strongest performance, consistently outperforming the other VLLMs. Furthermore, the open-source Llama demonstrates a significant performance gap compared to the closed-source VLLMs. We give some highlights associated with the benchmark results as follows:

Row-column Sensitivity Analysis. We found that all models show inconsistent performance between row/column-related tasks. To mitigate the influence of uneven distributions of rows and columns, leading to varying difficulty levels, we further refine the experimental results by selecting samples where the difference between the number of rows and columns does not exceed three. As shown in Figure 3, models perform better on column-related tasks, suggesting

Challenge	Scenario	Description
Visual Conditions	Blur	The image is out of focus, with details appearing smeared or indistinct.
	Underexposure	The image is too dark, which may cause the content to be unclear.
	Overexposure	The image is overly bright, losing detail in some regions.
Table Border Quality	Unclear Borders	The image's table borders are faint, blending into the background.
	Missing Borders	The table is without expected borders or separators.
	Thickened Borders	The table borders are thickened.
Geometric Deformation	Tilt 20°	The image is tilted at 20° angle.
	Tilt 40°	The image is tilted at 40° angle.

Table 2: Specific description of the challenges and scenarios in low-quality image inputs.

VLLMs tend to favor column-structured data. A likely reason is that, in many cases, columns represent diverse attributes while rows correspond to similar entities—making attribute-based (column) tasks more amenable to accurate recognition.

Image Quality Analysis. We conduct in-depth experiments and analysis on the TR task. The results demonstrate that the VLLMs perform relatively well on the SciTSR dataset with higher image quality using simple prompts. However, the performance gap is considerably more significant when processing the PubTabNet dataset with lower image quality, especially compared to traditional models. This phenomenon indicates that the quality of the input image is a key bottleneck limiting the performance of VLLMs. Section 3.4 provides a more in-depth analysis of this bottleneck.

3.4 Bottleneck Analysis

We further investigate the performance of VLLMs under varying image quality conditions and assess their visual robustness to these conditions through empirical analysis.

Experimental Setup

To comprehensively evaluate the visual robustness of VLLM, we focus on three distinct visual challenges: the image quality challenge, the table border quality challenge, and the geometric deformation challenge. Details of these challenges are provided in Table 2.

Visual Conditions. Visual conditions is a key factor affecting the accuracy of VLLMs in table recognition. To assess it, we systematically analyze the performance of VLLMs under various visual conditions across three scenarios: *blur*, *underexposure*, and *overexposure*. These analyses demonstrate the robustness of VLLMs in handling impaired visual conditions.

Table Border Quality. Table borders indicate structural information in table elements. We evaluate the impact of border visibility and completeness on table recognition performance by considering scenarios including *unclear table borders* and *missing borders*. Additionally, we explore the effect of border changes in the table by *thickened table borders*.

Geometric Deformation. Geometric deformation caused by viewing angles or operations can disrupt the geometric consistency of tables. We evaluated the robustness of VLLMs against geometric deformations by testing them *under tilt conditions of 20°* and *excessive tilt conditions of 40°*.

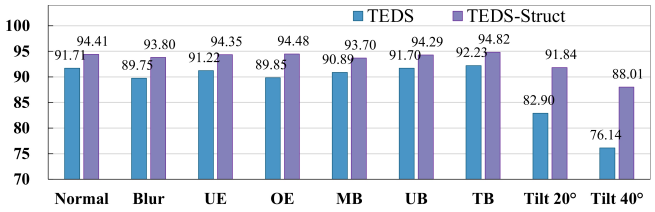

Figure 4: Evaluation results of bottleneck scenarios: abbreviations UE (Underexposure), OE (Overexposure), MB (Missing Borders), UB (Unclear Borders), TB (Thickened Borders).

Discussion and Analysis

As shown in Figure 4, while blurring and overexposure can degrade text clarity to some extent, the distortion caused by skewed tables severely disrupts structural integrity, significantly affecting the recognition performance of VLLM.

Although table borders are often considered essential for conveying structured information, fading or removing these borders has minimal impact on the performance of VLLMs. This result suggests that VLLMs do not heavily rely on borders. VLLMs only pay slight attention to the structural information the borders provide when thickening.

4 Neighbor-Guided Toolchain Reasoner

Through our in-depth analysis of VLLMs' performance on the benchmark, we have identified that improving input image quality is essential for enhancing VLLMs' capability to recognize and interpret structured image data more effectively. To address this, we propose the NGTR framework.

4.1 Framework Overview

NGTR enhances input image quality by applying various tool combinations tailored to low resolution, overexposure, and noise interference. As shown in Figure 5, we design a similarity-based neighbor retrieval module to select a suitable combination of tools. Subsequently, the tool invocation experience learning module executes each plan and generates the corresponding structured data to evaluate the effectiveness of different plans. Finally, we propose a reflection-driven tool selection module to integrate iterative tool invocation and dynamic feedback to refine the processing flow. The optimized image is then input into the VLLMs, which utilize their powerful reasoning capabilities to generate structured data.

4.2 Toolkit Preparation

Inspired by the conclusions of the Bottleneck Analysis (Section 3.4), we employ five distinct tools to address various scenarios and potential issues that may arise in the table recognition task. These tools are shown in Table 3. By combining different tools, NGTR effectively addresses various challenging situations. For example, when the table border is faint, the VLLMs can invoke the border enhancement tool to strengthen the structural information by thickening the table border.

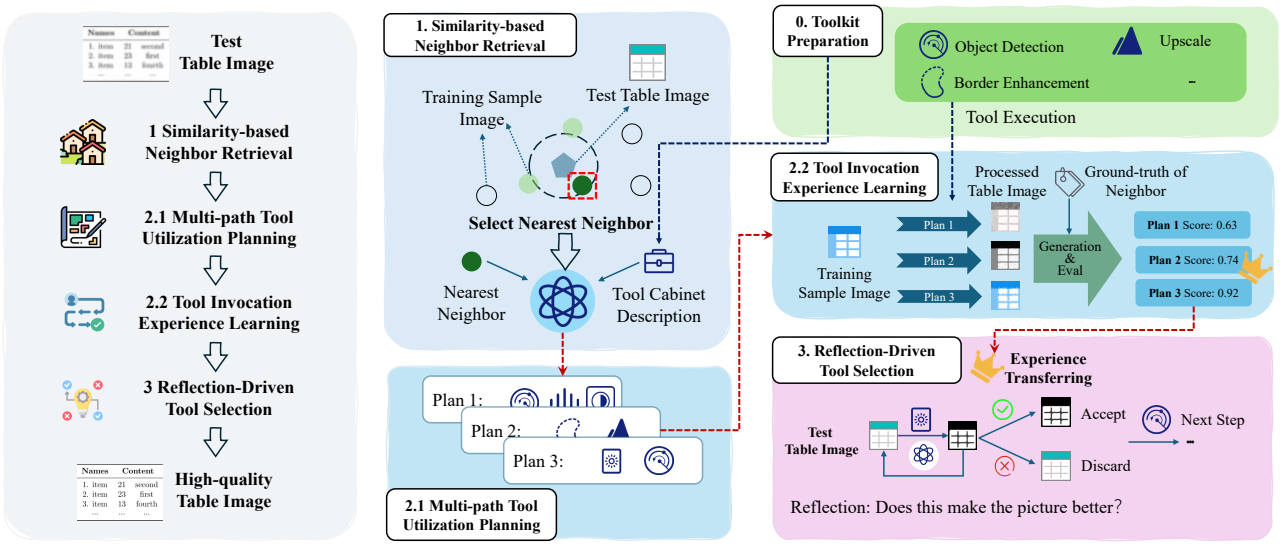

Figure 5: Illustration of a pipeline for table image preprocessing leveraging a toolkit of lightweight vision models.

Tool	Descriptions
Border Enhancement	The border enhancement tool improves the legibility of tables and their structures by thickening the border lines in the image. This process enhances the structural information features of tabular data.
Image Upscaling	Image upscaling optimizes image resolution to improve visual quality. This technique is commonly employed to repair and enhance blurry images.
Noise Reduction	The noise reduction tool enhances image quality by adjusting brightness and contrast to mitigate noise interference and underexposure issues.
Binarization	This tool converts images to black and white, highlighting key features for easier extraction.
Detection and Cropping	This tool identifies table regions within an image and crops them into independent segments. It is particularly suitable for processing images of tables embedded in complex backgrounds.

Table 3: Specific descriptions of built tools in the toolkit.

4.3 Similarity-based Neighbor Retrieval

Neighbor retrieval methods enable VLLMs to retrieve similar neighbor samples, providing richer contextual information. In the NGTR framework, we hypothesize that images with similar features exhibit similar results after being processed by the same image preprocessing toolchain. Consequently, the processing results of neighbor samples could guide the selection of a potentially optimal toolchain for test samples. We first retrieve images that are similar to the target task images from a sample set. Subsequently, we employ a prompting template to leverage the VLLM's planning capabilities to generate tool invocation plans. The retrieval process can be formally described as follows:

$$\text{Retrieval}(I^{\text{test}}, \mathcal{D}', f) = \arg\max \left[f\left(I^{\text{test}}, I^i\right) \right]_{i=1}^{|\mathcal{D}'|}, \quad (1)$$

where I^{test} represents an image from the test set, $\mathcal{D}'$ denotes a subset of the training dataset, and f is the similarity measurement function. In this paper, we combine the ORB (Oriented FAST and Rotated BRIEF) algorithm with the Hamming distance as f to measure the similarity between images. Then, we guide the VLLMs to generate multiple tool invocation plans for the image. The generation process can be formally expressed as follows:

$$\text{VLLM}(\mathcal{T}, \mathcal{N}(I^{\text{test}})) \rightarrow \{p_1, p_2, \ldots, p_n\}, \quad (2)$$

where $\mathcal{T}$ represents the description information set of all available image preprocessing tools, including their functions, applicable scenarios, invocation identifiers, and other relevant details; $\mathcal{N}(I^{\text{test}})$ denotes the neighbor image samples of the test sample I^{test}, along with their associated features, retrieved from the training set; and $\{p_1, p_2, \ldots, p_n\}$ represents the generated candidate set of plans, which are then used to select an appropriate tool invocation plan.

4.4 Tool Invocation Experience Learning

In this module, we follow a sequential workflow to evaluate the multiple tool invocation plans. First, we execute each tool invocation plan generated by the previous module to obtain multiple processed images:

$$I_{p_i} = f_{p_i}(I), \quad i \in \{1, 2, \ldots, n\}, \quad (3)$$

where f_{p_i} denotes the image preprocessing tools. Next, we employ a prompt template to guide the VLLMs in generating a markup sequence. Subsequently, we evaluate the prediction results based on the example labels. The evaluation process employs the tree edit distance-based similarity (TEDS) metrics to quantify the accuracy of the VLLMs output. By following this process, we calculate a quantitative score for each toolchain, enabling the selection of a suitable plan.

4.5 Reflection-driven Tool Selection

Although the tool invocation experience learning module provides a high-quality plan, mindlessly applying the tool invocation plan to new samples may result in the loss of critical information in the image, thereby affecting the accuracy of

Dataset	Metrics	Lightweight OCR Model			Prompt Tuning in VLLMs									
		EDD	LGPMA	LORE	Phi	GPT-mini	Qwen	Llama	Gemini			GPT-4o		
									direct	NGTR	Δ	direct	NGTR	Δ
SciTSR	TEDS	-	**95.08**	-	66.18	87.18	89.40	87.24	90.15	91.07	+0.92	90.70	**92.58**	+1.88
	TEDS-Struct	-	96.24	**97.22**	71.56	92.03	93.06	92.31	93.73	95.09	+1.36	94.20	**95.78**	+1.58
PubTabNet	TEDS	89.67	**94.63**	-	49.92	58.68	52.53	79.04	81.00	84.80	+3.80	74.46	**85.03**	+10.57
	TEDS-Struct	-	96.70	**96.94**	57.65	73.00	63.90	87.64	85.28	89.30	+4.02	84.91	**92.31**	+7.40
WTW	TEDS-Struct	-	-	**93.86**	-	31.72	-	32.87	42.62	44.68	+2.06	40.01	**52.03**	+12.02

Table 4: Performance comparison of methods on the SciTSR, PubTabNet, and WTW datasets. "-" indicates the method's lack of results (specific reasons are provided in the implementation details). Best scores in the lightweight OCR model category are highlighted in blue, while best scores in the prompt tuning in VLLMs category are highlighted in green.

the final result. To address this, we introduce the reflection-driven tool selection module during the execution phase to refine the processing flow, reduce information loss, and thereby improve recognition accuracy. The formalized expression of the reflection module is as follows:

Let $I^{(t-1)}$ denote the image before the t-th operation and $I^{(t)}$ denote the image after the t-th operation. The VLLMs computes $\gamma^{(t)}$ to determine whether to accept the operation:

$$\gamma^{(t)} = \text{reflect}(I^{(t-1)}, I^{(t)}), \tag{4}$$

where $\gamma^{(t)}$ is a binary decision indicating the quality change between the before and after images. The function reflect$(\cdot)$ evaluates the difference in quality. If $\gamma^{(t)} = 1$, the operation is considered successful; otherwise, if $\gamma^{(t)} = 0$, the operation is rejected, and the process proceeds to the next step.

This step-by-step module enhances the interaction between the VLLMs and the target image, ensuring the accuracy of the final task outcome. More importantly, introducing this module enables downstream researchers and developers to flexibly customize and expand the toolkit without worrying about the impact of poorly performing expanded tools on the final results, thereby significantly improving the versatility and transferability of our framework. In the last step, we design a simple prompt template to instruct VLLM to generate a markup sequence and obtain the result of table recognition.

5 Experiments

5.1 Experimental Setup

Datasets

In this study, we utilize three widely-used table recognition datasets: SciTSR [Chi *et al.*, 2019], PubTabNet [Zhong *et al.*, 2020], and WTW [Long *et al.*, 2021], each offering unique characteristics and challenges. SciTSR is a dataset comprising tables extracted from the scientific literature, and the image quality in this dataset is relatively high. In contrast, the image resolution of PubTabNet is 72 pixels per inch, and its overall image quality is relatively low. WTW contains images collected from the wild, introducing a variety of extreme cases, such as tilt, blur, and table curvature. These datasets encompass diverse table types and various unique visual challenges, providing a robust foundation for benchmarking.

Baselines

We select six Vision Large Language Models (VLLMs), including Phi[1], Llama[2], GPT-mini, Qwen[3], GPT[4], and Gemini[5], as baseline models for comparison. Additionally, we select three representative deep learning-based methods as baselines for comparison: EDD [Zhong *et al.*, 2020] based on sequence modeling, LGPMA [Qiao *et al.*, 2021] based on cell bounding box detection, and LORE [Xing *et al.*, 2023] based on cell point center detection.

Evaluation Metrics

As for the evaluation metrics of TR, we use a similarity metric based on Tree-Edit Distance (TEDS) [Zhong *et al.*, 2020] and the TEDS-Struct metric. We employ two evaluation metrics for the hierarchical tasks described in Section 3.2: accuracy (ACC) and micro-averaged F1 score (F1-score). Specifically, ACC is used to evaluate cell-level tasks (excluding merged cell detection) and table-level tasks; the F1-score is utilized for row- and column-level tasks and merged cell detection.

Implementation Details

For the PubTabNet dataset, We randomly select 1,500 images from the validation set. For the SciTSR and WTW datasets, we use their complete test sets for evaluation. Since the WTW dataset does not provide content information for table recognition, we do not report its TEDS scores.

For LORE, since it is mainly aimed at table structure recognition but not table content recognition, we only report its performance scores for table structure recognition. As for EDD, since its model training requires a large amount of end-to-end annotated data, and SciTSR and WTW lack corresponding labeled data, its performance on these datasets has not been evaluated. LGPMA depends on table content for training, but since WTW lacks content labels, its performance on this dataset was not assessed.

[1]https://azure.microsoft.com/en-us/products/phi/
[2]https://www.llama.com/
[3]https://qwenlm.github.io/blog/qwen-vl/
[4]https://openai.com/index/hello-gpt-4o/
[5]https://deepmind.google/technologies/gemini/pro/

Method	SciTSR		PubTabNet	
	TEDS	TEDS-Struct	TEDS	TEDS-Struct
NGTR	92.56	95.43	85.03	92.31
w/o EXP	90.33	93.68	80.57	88.40
w/o REF	91.53	94.77	82.08	91.85

Table 5: Ablation study results on key components of the framework. "EXP" denotes the tool invocation experience learning module, "REF" represents the reflection-driven tool selection module.

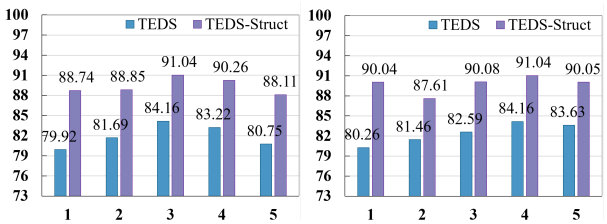

Figure 6: The impact of the number of tools (shown on the left) and the number of multi-paths (shown on the right) on performance.

5.2 Main Results Analysis

Tables 4 show our benchmark results on the table recognition and table structure recognition tasks. Based on the experimental results, we draw the following insights:

Performance Analysis of NGTR

As shown in Table 4, the experimental results compare our NGTR framework with baseline methods. The main results show that our framework achieves significant performance improvements on the PubTabNet dataset, mainly attributed to our framework's enhanced VLLMs robustness when dealing with low-quality inputs. On the SciTSR dataset, our framework also outperforms all VLLMs baselines, further verifying our framework's effectiveness.

Open-source and Closed-source VLLMs

The advanced open-source VLLMs demonstrate capabilities comparable to closed-source VLLMs in this task. As a representative of open-source models, Llama exhibits outstanding performance, particularly in the table structure recognition task (TEDS-Struct) on the PubTabNet dataset. Llama achieves a relatively better result, surpassing GPT by 2.73 points and Gemini by 2.36 points. These results confirm the potential of open-source VLLMs for further research.

WTW Dataset: A Challenge for VLLMs

Table 4 presents the experimental results on the WTW dataset. These results indicate that prompt-tuning based VLLMs methods still exhibit significant gaps compared to traditional lightweight OCR methods, highlighting the challenges VLLMs face when processing datasets with wild scenarios. Further analysis of the outputs suggests that performance declines notably when VLLMs process tables with numerous empty cells, unevenly distributed text, and skewed, rotated, or densely packed text. Further analysis of the outputs suggests that VLLMs tend to ignore cells lacking semantic content. While this behavior helps avoid processing irrelevant data, it also limits their ability to effectively capture the structural information of tables with many blank cells.

5.3 Ablation Study w.r.t Key Components

Effectiveness of Tool Invocation Experience Learning. We performed an ablation study by removing the tool invocation experience learning module. In this experiment, we only led the VLLMs to generate a tool invocation plan and applied it directly to the test samples. As shown in Table 5, while the VLLMs could generate a valid tool invocation plan, the lack of effective validation of the generated plan led to a significant performance decline. This further validates the critical role of the tool invocation experience learning module in improving the NGTR framework performance.

Effectiveness of Reflection-driven Tool Selection. We performed an ablation study by removing the reflection-driven tool selection module. We apply all the tools to the task images in a single pass. The results are presented in Table 5, without the reflection module for stepwise backtracking validation, VLLMs cannot effectively supervise the processing procedure, which may lead to the incorporation of unsuitable tools for the current sample, thereby affecting performance.

5.4 Hyperparameter Sensitivity Analysis

The NGTR framework contains two core parameters: the maximum length of the toolchain execution plan L and the number of plans generated each time N. As shown in Figure 6, a moderate toolchain length achieves an adequate balance between complexity and performance, as excessive toolchain length increases combinatorial complexity and limits processing performance, thereby affecting the framework's ability to generate high-quality solutions. Similarly, generating a moderate number of execution plans effectively balances solution quality and generation efficiency, whereas generating too few or too many plans slightly reduces performance. Therefore, a moderate toolchain length and number of execution plans can balance complexity and performance well, providing valuable guidance for the tool invocation.

6 Conclusion and Limitation

This paper addressed the under-explored challenge of table recognition using VLLMs in a training-free paradigm. We proposed the NGTR framework, which enhanced input image quality through lightweight models and neighbor-guided tool invocation strategies. Extensive experiments demonstrated that NGTR significantly improved VLLM-based table recognition performance. This work not only established a benchmark for table recognition but also highlighted the potential of VLLMs in advancing table understanding, paving the way for future research and applications.

Limitation. Despite the strengths of our framework, we acknowledge several limitations that warrant further investigation. Firstly, its performance depends on the underlying toolkit. Secondly, when the available set of neighbor candidates does not sufficiently cover a wide range of scenarios, selecting an inappropriate neighbor may lead to suboptimal performance. Nevertheless, we believe the NGTR framework demonstrates strong generalizability, serving as a versatile approach for tool invocation for various domains.

Acknowledgements

This research was supported by grants from the grants of Provincial Natural Science Foundation of Anhui Province (No.2408085QF193), USTC Research Funds of the Double First-Class Initiative (No. YD2150002501), the National Key Research and Development Program of China (Grant No. 2024YFC3308200), the National Natural Science Foundation of China (62337001), the Key Technologies R & D Program of Anhui Province (No. 202423k09020039) and the Fundamental Research Funds for the Central Universities (No. WK2150110032).

References

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, et al. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, et al. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chang *et al.*, 2024] Yupeng Chang, Xu Wang, Jindong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[Chen *et al.*, 2023] Leiyuan Chen, Chengsong Huang, Xiaoqing Zheng, et al. Tablevlm: Multi-modal pre-training for table structure recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 2437–2449, 2023.

[Chen *et al.*, 2024] Shijie Chen, Yu Zhang, and Qiang Yang. Multi-task learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32, 2024.

[Cheng *et al.*, 2025a] Mingyue Cheng, Yucong Luo, et al. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*, 2025.

[Cheng *et al.*, 2025b] Mingyue Cheng, Qingyang Mao, et al. A survey on table mining with large language models: Challenges, advancements and prospects. *Authorea Preprints*, 2025.

[Chi *et al.*, 2019] Zewen Chi, Heyan Huang, Heng-Da Xu, et al. Complicated table structure recognition. *arXiv preprint arXiv:1908.04729*, 2019.

[Deng *et al.*, 2022] Xiang Deng, Huan Sun, Alyssa Lees, et al. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

[Gou *et al.*, 2023] Zhibin Gou, Zhihong Shao, Yeyun Gong, et al. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

[Guan *et al.*, 2023] Lin Guan, Karthik Valmeekam, Sarath Sreedharan, et al. Leveraging pre-trained large language models to construct and utilize world models for model-based task planning. *Advances in Neural Information Processing Systems*, 36:79081–79094, 2023.

[Guo *et al.*, 2023] Jiaxian Guo, Junnan Li, Dongxu Li, et al. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023.

[Guo *et al.*, 2024] Zhixin Guo, Mingxuan Yan, Jiexing Qi, et al. Adapting knowledge for few-shot table-to-text generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[Hu *et al.*, 2024] Anwen Hu, Haiyang Xu, Jiabo Ye, et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024.

[Huang *et al.*, 2023] Yongshuai Huang, Ning Lu, Dapeng Chen, et al. Improving table structure recognition with visual-alignment sequential coordinate modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11134–11143, 2023.

[Kieninger and Dengel, 1999] Thomas Kieninger and Andreas Dengel. The t-recs table recognition and analysis system. In *Document Analysis Systems: Theory and Practice (DAS'98 Proceedings)*, pages 255–270. Springer, 1999.

[Kojima *et al.*, 2022] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, et al. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[Li *et al.*, 2024] Junyi Li, Tianyi Tang, Wayne Xin Zhao, et al. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39, 2024.

[Liu *et al.*, 2024a] Yuliang Liu, Biao Yang, et al. Textmonkey: An ocr-free large multimodal model for understanding document. *arXiv preprint arXiv:2403.04473*, 2024.

[Liu *et al.*, 2024b] Zhiding Liu, Jiqian Yang, Mingyue Cheng, et al. Generative pretrained hierarchical transformer for time series forecasting. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2003–2013, 2024.

[Long *et al.*, 2021] Rujiao Long, Wen Wang, et al. Parsing table structures in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 944–952, 2021.

[Luo *et al.*, 2024] Yucong Luo, Qitao Qin, Hao Zhang, et al. Molar: Multimodal llms with collaborative filtering alignment for enhanced sequential recommendation. *arXiv preprint arXiv:2412.18176*, 2024.

[Ma *et al.*, 2023] Chixiang Ma, Weihong Lin, Lei Sun, et al. Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognition*, 133:109006, 2023.

[Mao *et al.*, 2024] Qingyang Mao, Qi Liu, Zhi Li, et al. Potable: Programming standardly on table-based reasoning like a human analyst. *arXiv preprint arXiv:2412.04272*, 2024.

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

[Qiao *et al.*, 2021] Liang Qiao, Zaisheng Li, Zhanzhan Cheng, et al. Lgpma: Complicated table structure recognition with local and global pyramid mask alignment. In *International conference on document analysis and recognition*, pages 99–114. Springer, 2021.

[Qin *et al.*, 2024] Chunxia Qin, Zhenrong Zhang, et al. Semv3: A fast and robust approach to table separation line detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

[Ren *et al.*, 2016] Shaoqing Ren, Kaiming He, Ross Girshick, et al. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[Salaheldin Kasem *et al.*, 2024] Mahmoud Salaheldin Kasem, Abdelrahman Abdallah, Alexander Berendeyev, et al. Deep learning for table detection and structure recognition: A survey. *ACM Computing Surveys*, 56(12):1–41, 2024.

[Schreiber *et al.*, 2017] Sebastian Schreiber, Stefan Agne, Ivo Wolf, et al. Deepdesrt: Deep learning for detection and structure recognition of tables in document images. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1. IEEE, 2017.

[Shigarov *et al.*, 2016] Alexey Shigarov, Andrey Mikhailov, and Andrey Altaev. Configurable table structure recognition in untagged pdf documents. In *Proceedings of the 2016 ACM symposium on document engineering*, pages 119–122, 2016.

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

[Siddiqui *et al.*, 2019] Shoaib Ahmed Siddiqui, Imran Ali Fateh, Syed Tahseen Raza Rizvi, et al. Deeptabstr: Deep learning based table structure recognition. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 1403–1409. IEEE, 2019.

[Sui *et al.*, 2024] Yuan Sui, Mengyu Zhou, et al. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 645–654, 2024.

[Wan *et al.*, 2024] Jianqiang Wan, Sibo Song, Wenwen Yu, et al. Omniparser: A unified framework for text spotting key information extraction and table recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15641–15653, 2024.

[Wang *et al.*, 2011] Kai Wang, Boris Babenko, and Serge Belongie. End-to-end scene text recognition. In *2011 International conference on computer vision*, pages 1457–1464. IEEE, 2011.

[Wang *et al.*, 2024a] Jiahao Wang, Mingyue Cheng, et al. Tabletime: Reformulating time series classification as training-free table understanding with large language models. *arXiv preprint arXiv:2411.15737*, 2024.

[Wang *et al.*, 2024b] Zilong Wang, Hao Zhang, Chun-Liang Li, et al. Chain-of-table: Evolving tables in the reasoning chain for table understanding. *arXiv preprint arXiv:2401.04398*, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Xing *et al.*, 2023] Hangdi Xing, Feiyu Gao, et al. Lore: logical location regression network for table structure recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2992–3000, 2023.

[Ye *et al.*, 2023] Jiabo Ye, Anwen Hu, Haiyang Xu, et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2841–2858, 2023.

[Ye *et al.*, 2024] Han-Jia Ye, Si-Yang Liu, Hao-Run Cai, et al. A closer look at deep learning on tabular data. *arXiv preprint arXiv:2407.00956*, 2024.

[Yin *et al.*, 2024] Shukang Yin, Chaoyou Fu, Sirui Zhao, et al. A survey on multimodal large language models. *National Science Review*, page nwae403, 2024.

[Zhang *et al.*, 2024] Yiming Zhang, Yaping Zhang, Lu Xiang, and Yu Zhou. Multi-modal attention based on 2d structured sequence for table recognition. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 378–391. Springer, 2024.

[Zhao *et al.*, 2024] Weichao Zhao, Hao Feng, Qi Liu, et al. Tabpedia: Towards comprehensive visual table understanding with concept synergy. *arXiv preprint arXiv:2406.01326*, 2024.

[Zheng *et al.*, 2021] Xinyi Zheng, Douglas Burdick, et al. Global table extractor (gte): A framework for joint table identification and cell structure recognition using visual context. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021.

[Zheng *et al.*, 2024] Mingyu Zheng, Xinwei Feng, Qingyi Si, et al. Multimodal table understanding, 2024.

[Zhong *et al.*, 2020] Xu Zhong, Elaheh ShafieiBavani, and Antonio Jimeno Yepes. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pages 564–580. Springer, 2020.