Multimodal Tabular Data Learning

Jun-Peng Jiang

School of Artificial Intelligence, Nanjing University, Nanjing, China
State Key Laboratory for Novel Software Technology, Nanjing University, China
jiangjp@lamda.nju.edu.cn

Abstract

Tabular data is a fundamental form of information in real-world applications, ranging from finance and healthcare to scientific research. Unlike traditional views that treat tables as isolated structured data, tables are often inherently multimodal—appearing as images, embedded in documents, or coexisting with text and other modalities. My research explores multimodal tabular data learning, aiming to bridge structured tabular knowledge with diverse input forms and tasks. To this end, our work investigates leveraging tabular data as expert knowledge to provide guidance for visual modalities and enable cross-modal transfer learning. We also study more common scenarios where tables appear as images, conducting comprehensive investigations from evaluation to method development for table-based question answering and reasoning. Beyond these works, we extend tabular learning to more general scenarios, developing unified models capable of handling diverse table tasks within a single framework, and further expanding from tables to broader document-level parsing and understanding.

Introduction

Tabular data, with its structured, row-and-column organization, is one of the most widely used data types in machine learning (Ye et al. 2024; Liu et al. 2024; Jiang et al. 2025a). Tables appear in numerous domains, including finance, healthcare, scientific research, and government statistics, where they serve as concise and interpretable representations of complex information (Gupta et al. 2022). While conventional approaches often treat tables as purely structured data, many real-world tables are inherently multimodal—they may appear as images, be embedded in documents, or coexist with text and other contextual information. This multimodality introduces unique challenges for machine learning models, which must not only interpret the table structure but also integrate information across different input modalities.

In practice, tables support a wide range of applications, but these uses also raise several interconnected research challenges. First, tables in multimodal forms require models to accurately interpret their structure and content. Second, tables contain rich, structured knowledge that can guide learning in other modalities, motivating research on effective cross-modal knowledge transfer. Third, tables support diverse downstream tasks, including question answering, prediction, anomaly detection, and generation, which calls for unified and generalizable models that can extend from tables to broader document-level parsing. These challenges form a natural progression: *understanding tables in multimodal contexts, leveraging their knowledge across modalities, and supporting a wide range of table-related applications*.

Focusing on understanding tables in multimodal contexts, we address both evaluation and modeling challenges (Jiang et al. 2025c,b). From an evaluation perspective, we introduce the Massive Multimodal Tabular Understanding (MMTU) benchmark, which covers 8,921 QA pairs across multiple domains and assesses element-level understanding, row/column reasoning, compositional conditions, and basic calculations. From a modeling perspective, we enhance multimodal large language models with structure-aware visual encoders and privileged structured information to better capture table structure, align visual and textual inputs, and perform multi-step reasoning. These efforts provide both rigorous evaluation tools and effective model designs for advancing table understanding in multimodal scenarios.

To leverage tabular knowledge across modalities, we study how structured tables can provide expert guidance for other data types, particularly images (Jiang et al. 2024). Real-world challenges include accurately mapping diverse numerical and categorical tabular attributes to visual contexts and identifying which features are relevant for transfer. To address this, we develop techniques that align tabular attributes with image channels, enabling selective knowledge transfer that enhances visual prediction. By capturing relationships between numerical and categorical features and visual representations, this approach improves both the accuracy and interpretability of image classifiers, demonstrating the potential of tables as a source of structured knowledge for cross-modal learning.

Finally, to support a wide range of table-related applications, we explore unified models that can handle diverse table-related tasks, including question answering, prediction, generation, and anomaly detection, and further extend these models to document-level parsing. This direction aims to provide a flexible framework for comprehensive tabular and multimodal understanding, which we will describe in

more detail in subsequent sections.

General Tabular Applications

Unified Tabular Model

With the rapid development of large language models (LLMs), many natural language processing tasks, such as captioning, question answering, and summarization, have been successfully unified under a single model framework. In contrast, tables present unique challenges that make unification more difficult. Tables are inherently heterogeneous, combining numerical, categorical, and textual data, and often feature complex row-column structures, hierarchical relationships, and long sequences of entries. They may also contain missing or inconsistent values, which further complicates reasoning and prediction. Developing unified models that can handle diverse table tasks—including question answering, prediction, generation, and anomaly detection—is therefore both practically important and scientifically meaningful.

Despite this potential, several challenges arise from the nature of tables themselves. First, the heterogeneous and structured nature of tables requires models to understand both content and layout, which is more complex than processing linear text. Second, tables can be extremely large or deeply nested, making it difficult for standard LLMs to encode and reason over all relevant information. Third, subtle semantic distinctions between similar entries, or dependencies across rows and columns, demand fine-grained understanding that typical LLMs may not naturally capture. Addressing these challenges is essential for building generalizable models capable of performing multiple table tasks effectively.

To address the challenges of heterogeneous and complex tables, we adopt an agent-based approach for general tabular understanding. In this framework, the LLM serves as a central reasoning "brain," orchestrating task execution while delegating specific subtasks to specialized tools that are better suited for them. For instance, different components may handle numerical computation, categorical reasoning, or table layout interpretation, allowing the system to leverage the strengths of each module. This design enables flexible, modular, and interpretable processing of diverse table tasks, ensuring that the model can effectively perform question answering, prediction, generation, and anomaly detection across a wide range of table types and sizes. By structuring the model in this way, we aim to combine the general reasoning capability of LLMs with task-specific expertise, improving both performance and robustness in real-world applications.

Extend Tables to Documents

In real-world scenarios, tables often appear embedded within various types of documents, such as reports, articles, and forms. Understanding these tables requires not only interpreting the table structure itself but also parsing surrounding textual context, as both contribute to accurate comprehension. Accurately recognizing the content and layout of tables is therefore a critical first step toward effective document-level table understanding.

However, several challenges make this task difficult. Tables are highly heterogeneous and may include merged cells, hierarchical structures, or irregular layouts. In addition, the natural reading order of the document—including how text and tables are interleaved—can significantly affect interpretation. To address these challenges, we design a synthetic data engine that generates high-quality artificial data mimicking the distributions and structures of real-world documents. This approach allows models to learn robust parsing strategies across diverse layouts and content types, providing a foundation for accurate table and document understanding without relying exclusively on costly human-labeled data.

Conclusion

In this thesis, I focus on Multimodal Tabular Data Learning, exploring how structured tabular knowledge can be understood, leveraged across modalities, and generalized to multiple tasks. As of the submission date, the work described in the Introduction—covering the understanding of tables in multimodal contexts, cross-modal knowledge transfer, and preliminary steps toward generalizable systems—has been independently completed. The research presented in the General Tabular Applications section, including the development of unified models for diverse table tasks and the extension to document-level parsing, represents anticipated progress that I plan to complete by the workshop date. Together, these efforts aim to advance the capabilities of AI systems in interpreting, reasoning over, and generalizing from multimodal tabular data in real-world applications.

References

Gupta, V.; Zhang, S.; Vempala, A.; He, Y.; Choji, T.; and Srikumar, V. 2022. Right for the right reason: Evidence extraction for trustworthy tabular reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3268–3283.

Jiang, J.-P.; Liu, S.-Y.; Cai, H.-R.; Zhou, Q.; and Ye, H.-J. 2025a. Representation learning for tabular data: A comprehensive survey. *CoRR*, abs/2504.16109.

Jiang, J.-P.; Xia, Y.; Sun, H.-L.; Lu, S.; Chen, Q.-G.; Luo, W.; Zhang, K.; Zhan, D.-C.; and Ye, H.-J. 2025b. Multimodal Tabular Reasoning with Privileged Structured Information. In *NeurIPS*.

Jiang, J.-P.; Ye, H.-J.; Wang, L.; Yang, Y.; Jiang, Y.; and Zhan, D.-C. 2024. Tabular Insights, Visual Impacts: Transferring Expertise from Tables to Images. In *ICML*.

Jiang, J.-P.; Zhou, T.; Zhan, D.-C.; and Ye, H.-J. 2025c. Compositional Condition Question Answering in Tabular Understanding. In *ICML*.

Liu, S.-Y.; Cai, H.-R.; Zhou, Q.-L.; and Ye, H.-J. 2024. TALENT: A Tabular Analytics and Learning Toolbox. *CoRR*, abs/2407.04057.

Ye, H.-J.; Liu, S.-Y.; Cai, H.-R.; Zhou, Q.-L.; and Zhan, D.-C. 2024. A closer look at deep learning on tabular data. *CoRR*, abs/2407.00956.