OCR Table Extraction and Compliance Verification of Legal Documents Based on Multimodal Transformer

WeiZhong Tang
Department of Smart Policing Technology, Beijing Police College, Beijing 102202, China
Corresponding author: WeiZhong Tang (e-mail: piratet@sina.com)

ABSTRACT Accurate extraction and intelligent interpretation of structured information from complex documents remain challenging due to heterogeneous data modalities, intricate layouts, and long-range semantic dependencies. This study proposes a multimodal document understanding framework integrating optical character recognition (OCR), multimodal Transformer encoding, graph neural networks (GNNs), and reinforcement-learning-based reasoning. A dedicated OCR and layout-analysis module is first employed to acquire textual, visual, and spatial information from complex document images. Subsequently, a multimodal Transformer encoder performs joint representation learning by fusing visual features, semantic content, and layout embeddings to achieve robust table structure reconstruction and structured information extraction. To model semantic relationships among extracted entities, a relational graph convolutional network is constructed for heterogeneous information representation and propagation. Furthermore, a rule-guided reinforcement learning mechanism is introduced to optimize reasoning paths and improve decision accuracy in complex verification tasks. Experimental results demonstrate that the proposed framework achieves a GriTS score of 0.927 for table reconstruction and an F1 score of 93.1% for end-to-end verification, outperforming existing document-understanding approaches while significantly improving processing efficiency. The proposed framework provides an effective methodology for multimodal information fusion, graph-based information propagation, intelligent signal representation, and adaptive reasoning in complex information-processing systems, offering potential applications in intelligent sensing, communication-oriented information analysis, and distributed decision-support environments.

INDEX TERMS multimodal information fusion, graph-based information propagation, intelligent signal representation, multimodal Transformer, graph neural network, adaptive reasoning

I. INTRODUCTION

MANY legal documents include very complicated tables, some having merged cells and containing both multipage spreads and nested structures as part of their design. Unfortunately, traditional OCR technology is not able to accurately reconstruct the logical structure and semantic information contained in those tables; inaccurate extraction of table data from legal documents can greatly diminish the accuracy and efficiency of any automated compliance verification that takes place after receipt of the documents. In general, existing systems typically rely on either one of two possible modalities (text or graphics) to perform verification, or use static rule-based reasoning to perform verification, making it difficult to manage complex entity-to-clause relationships or to handle dynamic compliance logic in legal documents. For these reasons, building an end-to-end system that is able to integrate multimodal data, accurately parse table structures, and is able to achieve interpretable compliance verification is of considerable research value and provides significant application potential.

In recent years, with the in-depth application of AI (Artificial Intelligence) technology in the field of document understanding, the automated processing of professional documents such as legal and financial documents has become a research hotspot. In particular, many important works have emerged in the fields of OCR, multimodal information fusion, structured information extraction and compliance verification, which provide theoretical basis and methodological reference for this paper [1,2]. In terms of OCR and table extraction of legal documents, early research was mostly based on traditional image processing methods and rule engines. For example, Chawla A [3] systematically reviewed character recognition and information retrieval technology in structured data extraction and pointed out the potential of multi-feature fusion in understanding complex layouts. For the table structure commonly found in legal documents, Samantapudi R K R [4] proposed a table

extraction method for financial and transaction documents, but did not fully handle complex situations such as merged cells and nested tables. Pappula K K [5,6] further explored the application of multimodal AI in document structured data extraction and confirmed the effectiveness of combining visual and text features. Based on Transformer, financial documents were classified, demonstrating the advantages of pre-trained architecture in domain adaptation. In the field of multimodal document understanding, with the popularization of the Transformer architecture, researchers have begun to focus on the joint modeling of visual, text, and layout information [7,8]. Polo L [9] studied the application of OCR and AI-based label verification system in food traceability, demonstrating the practicality of multimodal methods in real-world scenarios. Dutta S [10] introduced a framework known as Visformers, combining vision and transformer technologies for improved classification of complex documents; this framework provides an example of multimodal coding design and is a direct reference to multimodal coding design. Zhang Q [11] subsequently demonstrated the benefit of integrating cross-modal features in professional document processing by enhancing the process of feature fusion and employing transfer learning in the classification of multi-format government documents. Further more, Maldini A S [12] demonstrated the viability of end-to-end visual/text models in real-world applications utilizing Faster R-CNNs (Region-based Convolutional Neural Networks) and Tr-OCR (Optimized by Transformerbased Optical Character Recognition) for the detection of multimodal hidden advertisements. Bulatov K B [13] proposed a framework for the unified analysis of identity documents that was focused on layout analysis and information structuring as needed to process specialized documents. Mandvikar S [14] and Odofin T [15], respectively, explored the integration of large language models and multimodal transformers into intelligent document processing workflows and recommended their use as a basis for designing multimodal encoders addressed within this paper. While multimodal fusion and structured understanding have demonstrated po tential for legal document processing, end-to-end complex table parsing and dynamic compliance reasoning exhibit significant limitations at this time.

Traditional methods for legal entity extraction and compliance verification rely largely on rules and knowledge graphs. Naik V [16] evaluated named entity recognition techniques in legal text and identified some of their limitations when using plain text versus multimodal approaches. Sinha B [17] developed an end-to-end processing framework for legal documents to utilize large language models (LLMs) and OCR; however, its application was not adequate for processing complex tables and performing logical reasoning. Chen J [18] demonstrated how natural language processing can be used to interpret tax regulations within intelligent tax systems, thus emphasizing how integrating subject matter expertise into automated compliance processes is critical. Zhang T [19] developed a multimodal artificial intelligence framework enhanced with a knowledge graph for intelligent data integration and improvement of compliance with tax laws, demonstrating the importance of using structured representations of the knowledge base when performing complex compliance processes. Bezditnyi V [20] researched the application of artificial intelligence in tax compliance and compli ance to laws and regulations of the business world as being valuable in relation to the role of automated reviews within the fields of professional services. In addition, Parakala A [21] and Pingili R [22] explored the deployment and challenges of document understanding systems in actual compliance tasks from the perspec tives of government documents and banking and financial scenarios, respectively. In terms of reasoning and verification mechanisms, reinforcement learning and GNN have been gradually introduced to improve the flexibility and interpretability of the system. Talakola S [23] and Chansarkar A [24] respectively studied the application of AI in bill of lading image conversion and unstructured document extraction, reflecting the evolutionary trend from perception to cognition. However, existing research focuses on the information extraction stage and lacks explicit modeling and dynamic reasoning optimization of the complex relationship between the extracted entities and clauses, which is the core difficulty of legal compliance verification. The purpose of this paper is to build an end-to-end intelligent analysis system for legal documents to solve the key problems in complex table structure restoration and compliance verification. To achieve this goal,the main contributions of this paper are:1) introducing learnable modal perception bias in the multimodal encoder to explicitly control the interaction intensity between heterogeneous features;

2) designing a query-based table structure parsing mechanism to achieve accurate modeling of merged cells and nested structures; 3) applying a rule-guided reinforcement learning reasoning path optimization method to achieve interpretable dynamic compliance verification on the legal knowledge graph.

II. CONSTRUCTION OF MULTIMODAL LEGAL DOCUMENT AND COMPLIANCE VERIFICATION FRAMEWORK

A. INTEGRATION OF DEDICATED OCR AND LAYOUT ENGINES FOR LEGAL DOCUMENTS

The purpose of this section is to obtain high precision digitization and structured comprehension of the original document images as the base for an end-to-end legal document analysis process. The primary source used as input into this framework is a scanned image or PDF of a legal document which has a number of complex tables embedded within it. Existing advanced OCR engines can be integrated with the exception of those requiring specific font compatibility, layout format, and general noise in legal documents can improve accuracy of how well these tables are recognized due to the optimization capability of these two technologies [25,26]. In addition, the OCR engine can

output the content of all text characters found, including the absolute bounding box coordinates of each character on the page, as well as provide detailed layout analysis so that non-text elements can be distinguished from each other - e.g., paragraph/column text, table areas, title blocks, and seals [27]. The process also detects all areas of the document containing tables and extracts them as separate image objects, along with their positional data on the page of the entire document, which can then be used for future detailed analysis.

To allow the coordinate information of the text to be matched with visual features in future modules, the coordinates are brought to a uniform coordinate system. The intention of this is to ensure that any events, including misrecognition of characters or disorder of line order by OCR occur, are corrected through the validation and correction of key legal terms using regular expressions. Therefore, they can be treated in accordance with characteristics of language and layout rules of legal documents. The process of rearranging is done in accordance with reading order based on spatial coordinates. The output of this process is structured data, machine-readable, containing layout and coordinate information necessary for multiple understanding of the document to be completed. A diagram showing the data flow and module interdependencies between the source document and the final compliance result is given in Figure 1.

The multimodal framework for legal document analysis and compliance verification described in this paper is a cascaded pipeline structure, as seen in Figure 1. The starting point of this framework is the original input document. The input document is processed first using an OCR layout analysis engine (using both OCR and layout analysis) followed by a parallel processing of OCR and layout analysis output through two main branches located above and below the graph of the input document to produce the desired outputs of the framework. The upper branch is the multimodal table parsing network, which uses images from table regions along with corresponding OCR text and coordinates. After performing feature fusion/encoding via a multimodal transformer, the table parsing network outputs a serialized table structure through the use of a table structure parser. The lower branch is the compliance verification inference engine, which utilizes a graph network created from legal entity/clause extraction from full text (including tables) in order to make compliance verification inferences from the graph network using a reinforcement learning agent that is guided by "rules". The compliance verification inference engine produces output being the compliance verification conclusion as well as interpretable paths taken to achieve the compliance verification conclusion. A feedback loop is established between the table parsing module and the graph construction module. When table serialization fails to meet predefined structural confidence thresholds, the affected cells are flagged and their textual content is passed directly to the entity extraction process without structural guarantees. The graph network models uncertainty by assigning lower confidence weights to nodes derived from such flagged cells. This mechanism prevents single-point failures in table parsing from cascading into erroneous compliance conclusions.

B. DESIGN OF MULTIMODAL TRANSFORMER ENCODER INTEGRATING VISUAL, TEXTUAL, AND LAYOUT INFORMATION

In this part of this paper, an improved multimodal transformer encoder is designed to co encode the visual features of the table area image (hereinafter referred to as the table image), the semantic information of the text report and the spatial organization of the text in the table image, so as to provide a deeply integrated feature representation for table structure analysis [28,29]. ResNet (residual network) model based on Imagenet pre training can be used as the backbone model of visual modality to extract the grid feature map of table image (i.e. the visual feature of table image) [30,31]. The feature vector of each grid cell (from the extracted grid feature map) contains the visual texture, line texture and character shape information of the grid cell. The pretraining BERT (Bidirectional Encoder Representations from Transformers) model from the legal field can be used for text data, and the character sequence can be recognized as input from the text block of the table image through OCR. Next, the normalized center point coordinates, width, and height of each feature unit (image grid or text block) are calculated and mapped to a high-dimensional layout embedding vector through an independent two-layer feedforward network.

The fusion strategy of multimodal features is the core of this encoder design. Instead of simply concatenating the three features, they are treated as a heterogeneous feature sequence input to the Transformer encoder [32,33]. Specifically, the sequence is composed of the sum of image grid features, text block features, and their corresponding layout embeddings. To distinguish different modalities and promote multimodal interaction during the encoding process, a learnable modality type embedding vector is introduced into the input layer. In the encoder's self-attention mechanism, a modality-aware bias is further introduced as an additional term in the attention weight calculation, used to explicitly regulate the interaction intensity within the same modality and between features of different modalities.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B_{m(m')}\right)V \tag{1}$$

Q, K, V represent the query matrix, key matrix, and value matrix, respectively. These are obtained from the input feature sequence through linear transformation and are core components of the self-attention mechanism. d_k represents the dimension of the key vector, used to scale the dot product QK^T to avoid excessively large values that could cause the Softmax gradient to vanish. $B_{m(m')}$ is the modality-aware bias matrix. This modified multilayer Transformer encoder, through stacked self-attention

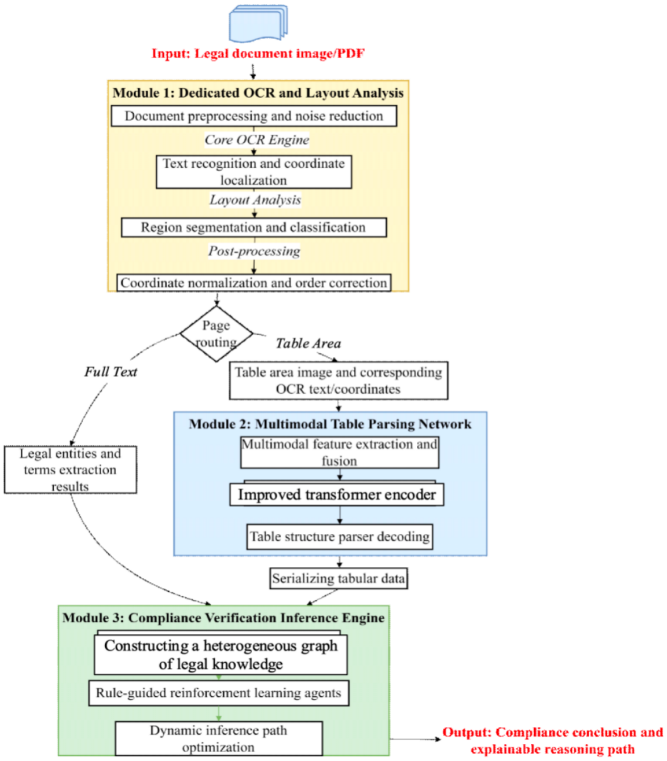

FIGURE 1. Data flow and module dependencies

and feedforward networks, iteratively fuses information from the three modalities, ultimately outputting a serialized feature representation containing rich visual-textual-layout contextual information. This representation is directly input into the downstream table structure parser.

C. COMPLEX TABLE STRUCTURE PARSER AND SERIALIZATION INFORMATION EXTRACTION

Based on the fused feature sequence output by the multimodal Transformer encoder, this section aims to design a table structure parser. The primary objective of the parser is to translate this combined data into an accurate logical representation of the table while producing structured data that can be read by machines. The parser has two primary objectives to accomplish at the same time: the first is to define the layout of the table's grid (i.e., row and column dividers) and how cells have been merged; the second is to link the text within each cell to the appropriate cell location.

To accomplish these goals, a query-based prediction method is used. The parser, while decoding, utilizes a set of learnable structural query vectors to determine the logical cell in the table using cross-attention computation with the encoder's multimodal feature sequences and iteratively collecting the visual, textual and layout context for each logical cell in the table during the decoding procedure.

After these steps have been completed for each query; each query result is passed on to one of three parallel prediction heads. The regression head provides bounding box coordinates for that logical cell in the normalized image coordinate system; the classification head provides the starting row, starting column, ending row, and ending column indices of that logical cell, which defines its row and column span; the relation head helps to identify all physical regions of a merged logical cell by calculating relevance scores between all of the queries to ensure structural consistency.

In the prediction process, this paper first consolidates all query results through integration, enabling them to be presented in a complete table structure. The last row and column of the table can be generated by taking all of the predicted horizontal bounding boxes and predicted vertical bounding boxes and clustering them. Each logical cell has its own unique identifier based on the range of rows and columns that it encompasses. The content of each logical cell consists of the OCR-identified text that is associated with it based on spatial inclusion rules. Each table is then serialized into a hierarchical data structure, which consists of an array of rows, along with row and column indices for each logical cell, each cell's merge-status and corresponding cell text content. This serialized structure maintains the logical two-dimensional relationships of the table's contents and can be directly computed as the data needed for verification compliance downstream. The parser receives a feature sequence from the multimodal encoder as input. Its core is a set of learnable cell queries that interact with the input features through a multi-layer decoder (based on a cross-attention mechanism). Each decoded query vector is fed into three prediction heads:

1) a bounding box regression head, which outputs the

precise coordinates of the cell. The bounding box regression loss uses the smoothed L1 loss function, defined as follows:

$$L_{bbox} = \sum_{i} smooth_{L1}(\hat{b}_i - b_i) \tag{2}$$

$\hat{b}_i$ is the predicted box coordinate, and b_i is the ground truth box coordinate.

2) Row and column index classification header, outputting the start row, end row, start column, and end column of the cell;

3) Cell relationship header, calculating the association score between query pairs, used for merging cell inference. The prediction results are integrated by the post-processing module to finally generate a serialized tabular data structure.

D. LEGAL ENTITY AND CLAUSE ASSOCIATION MODELING BASED ON GNN

From the serialized data obtained by the table parsing module, key legal entities and the specific legal clause numbers and contents referenced in the documents are extracted. To explicitly model the complex logical and constraint relationships between these discrete elements, this section constructs a heterogeneous legal knowledge graph and uses GNN for learning and representing the association relationships [34,35].

This heterogeneous graph contains two types of nodes: entity nodes and clause nodes. An entity type is defined by the text description, and it has a predetermined type. The node feature data is obtained from pretrained language model encodings of the text description. A clause node is based on one or more clauses of law, which are represented in the graph by their full text encoding. The creation of edges in the graph is based on several rules and legal principles as follows: if there is a direct reference or restriction relationship between entity node and clause node, each can be connected; if there is a strong semantic or logical relationship between two entity nodes, those are also connected. Each edge is assigned a specified relationship type. An R-GCN (Relational Graph Convolutional Network) is used to model this heterogeneous structure [36,37]. It enables the flow of information across node pairs via different edge types in a multi-layered manner by allowing multiple nodes to send and receive messages to other nodes at each layer. Each node collects the features from its neighbour nodes, with the amount of feature contribution from each neighbour determined by the type of edge connecting the two nodes and the features of the two nodes.

The update formula for node features in R-GCN is as follows:

$$h_v^{(l+1)} = \sigma\left(\sum_{r\in R}\sum_{u\in N_v^r}\frac{1}{|N_v^r|}W_r^{(l)}h_u^{(l)} + W_o^{(l)}h_v^{(l)}\right) \tag{3}$$

When an operation is performed on many levels multiple times, each node's final representational vector is able to hold a lot of contextual information about its local graph structure, like how to combine the meanings of other entities/terms related to it. After performing this series of operations on an entire graph, the graph becomes a set of node embeddings filled with relational semantic information. The resulting embeddings represent the semantic content of each individual node and preserve the logical dependency networks among all of the nodes in the original graph, giving machine-readable relational graphs to downstream agents that use rules to reason through the relationships between elements of the graph.

E. RULE-GUIDED REINFORCEMENT LEARNING COMPLIANCE REASONING PATH OPTIMIZATION

The centre of compliance verification has shifted from a traditional rule-related reasoning chain that provides very small and highly dynamic relationships among the many entities within a compliance graph network to a graph traversal and reasoning method that searches for compliance via graph nodes (i.e., entities) and edges (i.e., compliance constraints) in their associated legal provision code. To replace the existing static and weakly connected rule reasoning chain, this section provides a solution in the form of a rule-guided reinforcement learning agent, which allows it to learn and navigate through the graph network and learn the most effective sequence of compliance verification decisions within the graph.

An agent can be thought of as a moving through a decision making system within a graph. The current state of the agent is made up of the graph nodes that the agent is interested in now and the historical sequence of graph nodes that have been traversed by that particular agent up to now (referred to as the agent's "history"). The possible actions that the agent can take at any point in time are defined by all the nodes that are immediately adjacent (to the current node) and are connected to the current node via pre-defined relational edges (i.e., the next immediate nodes that can be traversed). The agent makes decisions based on a policy network that takes the current state of the agent as input, encodes that state into an embedding, and then provides the agent with a probability distribution on the set of all possible actions that the agent may take based on that embedding. Each time the agent makes a decision (i.e., selects an action) and moves to a new node, the environment provides the agent with a reward signal. The reward signal that the environment provides consists of two parts: a sparse rule-based reward, which is based on a rule set that is built in advance (or is an external resource); a dense exploration-based reward, which encourages the agent to traverse through shorter paths in order to verify the key clauses and interdependencies between entities, i.e. increase efficiency. The formal structure of the reward signal is given as follows:

$$R(s,a) = \lambda_1 \cdot R_{rule}(s,a) + \lambda_2 \cdot \left(-\frac{path\_len}{L_{max}} \right) + \lambda_3 \cdot R_{explore}(s) \tag{4}$$

To effectively inject domain knowledge into the learning process and improve early exploration efficiency, this paper introduces rule guidance into the agent's action selection mechanism. Specifically, a lightweight rule engine matches the type and attributes of the current node at each step, filtering out a logically consistent subset of candidate actions. The policy network primarily explores and utilizes this subset. This combination of "rule pruning" and "policy learning" ensures the logical foundation of the inference path while giving the model the flexibility to adapt to unforeseen complex situations. The rule engine operates as a logical filter rather than a decision maker, eliminating candidate nodes that violate explicit legal constraints while preserving all paths that satisfy basic consistency requirements. Reinforcement learning then selects among these admissible paths, optimizing for efficiency and coverage based on reward signals derived from successful verification examples. This separation of concerns allows the rule engine to remain lightweight by focusing only on necessary exclusions, while the RL agent captures complex patterns that are difficult to encode as static rules.

This design combines the deterministic nature of legal rules with the adaptive exploration capability of reinforcement learning. The rule engine prunes the action space by filtering candidate nodes that violate logical consistency, while the RL agent learns to select optimal paths within this constrained space, allowing the system to adapt to document-specific variations without sacrificing legal soundness. Through extensive interactive training with a simulated environment, the agent ultimately learns a policy that enables it to efficiently and accurately plan one or more verification paths to relevant clause nodes starting from any initial entity node, and outputs the decision basis on the path (i.e., the relationship between the visited node sequence and the trigger), thereby achieving interpretable compliance conclusion generation.

Figure 2 compares four path search strategies for compliance verification on graph networks. Among them, the "random walk" strategy serves as a performance lower bound; the "frequency-based heuristic search" simulates the baseline method based on keyword matching; the "GNN node importance-based search" utilizes the node centrality index learned by the GNN.

As shown in Figure 2, rule-guided reinforcement learning S4 provides better performance than the other strategies across all metrics measured. S4 achieves a verification accuracy of 93.0%, much higher than GNN node importance based search S3 (84.1%), word frequency based heuristic search S2 (75.6%), or random walk strategy S1 (58.2%). Additionally, the average length of the path found by S4 is 7.2 steps long, much shorter than any of the other strategies, meaning that S4 is able to find the key compliance nodes much more quickly than any of the other strategies. Overall, S4 covers 81.5% of the possible search space, which represents a high level of coverage while avoiding traversing any non-compliant nodes (invalid), which clearly demonstrates that S4 provides the best overall value compared to the other three strategies with respect to accuracy, efficiency, and total coverage. Therefore, it is demonstrated that the use of rule-guided and reinforcement learning mechanisms improves the compliance reasoning effect of legal documents.

III. EXPERIMENTAL VERIFICATION

A. EXPERIMENTAL SETUP AND DATASET

To verify the effectiveness of the constructed framework on legal document table understanding and compliance verification tasks, this paper uses the publicly available ICDAR 2019 cTDaR (Track B) and DocLayNet datasets. The former provides rich annotations of modern document table structures, suitable for training and evaluating table parsing models; the latter is a large-scale document layout analysis dataset, including fine-grained page segmentation and element classification in fields such as law and finance, providing an ideal foundation for multimodal feature fusion. The ICDAR 2019 cTDaR (Track B) dataset contains 600 training and 300 testing table images, with a substantial proportion of complex tables featuring merged cells and nested structures. The DocLayNet dataset comprises 8083 document pages across ten domains, including legal and financial documents; from this, 1500 legal-domain pages are selected for the compliance verification experiments, many of which correspond to real-world technical specifications and contracts.

Experiments are conducted on a server equipped with four NVIDIA A100 GPUs, implemented using PyTorch. Visual encoding uses ResNet-101, and text encoding uses a Chinese BERT model pre-trained on legal domain text. The table parser is configured with 128 learnable queries; the graph network uses a 3-layer R-GCN; the reinforcement learning agent is trained using the PPO (Proximal Policy Optimization) algorithm. Evaluation metrics include GridS and TEDS (Tree Edit Distance Similarity) for table structure reconstruction, and accuracy, recall, and F1 score for compliance verification.

B. PERFORMANCE EVALUATION OF TABLE STRUCTURE RECONSTRUCTION

This section evaluates a multimodal fusion framework for organizing complex tables in legal documents using the public test dataset ICDAR 2019 cTDaR (Track B). The evaluation compares the framework's performance with two most widely used table organization assessment features—GriTS and TEDS. Through these two widely adopted metrics, this paper measures the two restoration accuracies from two distinct dimensions: cell content accuracy and overall structural similarity.

This paper systematically compares the performance of four baseline models: (1) PaddleOCR v4, which is a

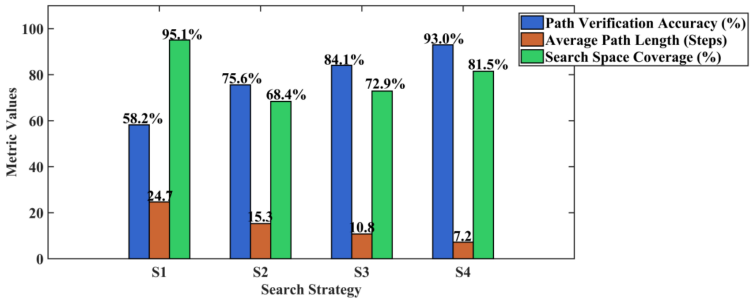

FIGURE 2. Performance comparison of different path search strategies on compliance verification tasks

commonly used open-source tool program for recognizing tables in industry; (2) A text-only baseline that reconstructs tables from the text sequence output by an OCR system using heuristic rules; (3) A visual-only baseline that uses a CNN-based semantic segmentation approach to predict the locations of table cells; and

(4) DocThinker, a general document understanding model that has been fine-tuned using the LayoutLMv3 architecture from the pre-trained weights available on Hugging Face. All comparisons are performed using the same dataset and hardware configuration. The quantitative results for the table structure reconstruction task (see Figure 3) indicate that the multimodal framework proposed in this paper performs best on both of the core metrics used to evaluate the models.

In Figure 3, Figure 3(a) and Figure 3(b) reveal that M1 achieves a GritS of 0.927, representing a 6% increase over PaddleOCR v4 (second best). The overall TEDS is 88.9. The framework achieves a GritS of 0.874 when processing complex tables typical of legal documents, with PaddleOCR v4 a GritS of 0.793. This demonstrates M1′s superior parsing ability by using a multimodal Transformer encoder and dedicated structure parser to combine visual, textual and layout data for building complex table layouts, including nested or merged tables. By providing a high accuracy mode of extracting structured information, M1 creates the reliable basis for performing compliance checks in the future.

To analyze the performance sources in more detail, this paper decomposes the reconstruction task into three subtasks for evaluation: cell location, row and column segmentation, and merged cell recognition (see Figure 4).

In Figure 4, the proposed framework comprehensively outperforms other models in cell location accuracy (0.941), row and column segmentation F1 score (0.953), and merged cell recognition F1 score (0.892). In particular, the F1 score for merged cell recognition is almost 16.6% greater than the next best performing model PaddleOCR V4 (0.765). The capabilities demonstrated are indicative of how well multimodal Transformer and query-based parsing mechanisms can model common merged and nested layouts throughout the legal document landscape. In contrast, text-only baselines or visual-only baselines perform extremely poorly on merged cell recognition. This highlights the challenges associated with relying solely on either modality (text or visual). While the general document model DocThinker (0.687) outperforms both the text-only or visual-only baselines, it does not provide performance comparable to that of specialized models developed for table parsing; therefore, a specialized fusion and parsing capability tailored to table elements is essential to provide the highest level of precision structured data for subsequent compliance verification activities.

Each test table sample is visualized based on its structural complexity (measured by the number of cells and the degree of merging) and the GriTS improvement of the proposed method relative to PaddleOCR v4, and the model error distribution is plotted, as shown in Figure 5.

Figure 5 uses table structure complexity (horizontal axis, range 0-1, 200 sample points generated from a random uniform distribution) as the independent variable and the GriTS improvement of the proposed method compared to PaddleOCR v4 as the dependent variable (vertical axis), plotting a scatter plot and overlaying a linear trend line. The data points change color from blue to red with complexity, showing that the improvement value is roughly distributed between 0 and 0.2, and exhibits a clear upward trend: the slope of the trend line is positive. This change indicates that the multimodal method applied in this paper has a stable gain when processing simple tables, while for complex tables (complexity > 0.4), its performance improvement is particularly significant, reaching over 0.15, strongly verifying that the multimodal framework integrating vision, text, and layout has a more prominent advantage in parsing complex legal tables, and its effectiveness further increases with the increase of table structure complexity.

C. END-TO-END COMPLIANCE VERIFICATION PERFORMANCE

Based on the precise parsing of document and table structures by the preceding modules, this section evaluates the end-to-end compliance verification performance of the framework on the publicly available DocLayNet dataset (results are shown in Figure 6). This dataset provides clear document layout segmentation and classification. Based on this, this paper simulates and constructs typical legal compliance verification tasks, involving six common problems such as “amount consistency” and “completeness of clause citation,” and generates corresponding truth labels

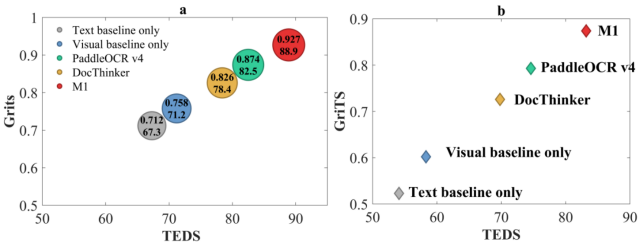

FIGURE 3. Comprehensive performance comparison analysis of table recognition models: (a) Overall performance comparison; (b) Performance comparsion of complex tables

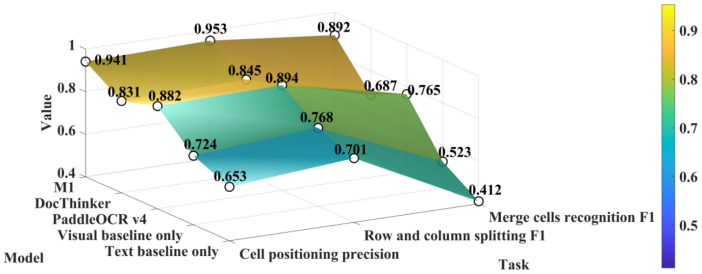

FIGURE 4. Performance comparison of various models in the table structure reconstruction subtask

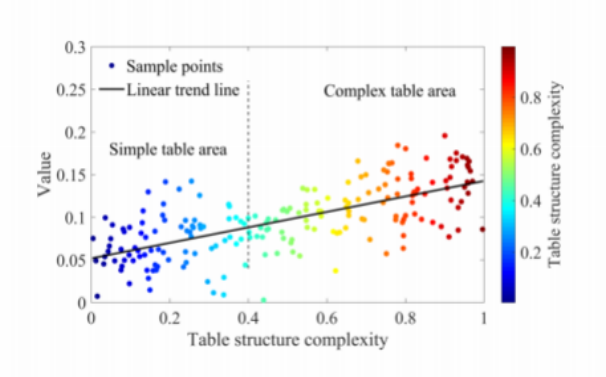

FIGURE 5. Relationship between performance improvement of the proposed method and table structure complexity

to ensure the objectivity and reproducibility of the evaluation.DocThinker, a multimodal document understanding model based on LayoutLMv3, is selected as the primary baseline for compliance verification. This model processes both text and layout information from document images, providing a direct comparison against a state-of-the-art multimodal approach.

To comprehensively evaluate the framework's performance in logical reasoning and decision-making, this study compares it with three benchmark frameworks: (1) Rule-Based Framework (RB), a rule-based approach utilizing multiple rule sets including text pattern matching rules (regular expressions) and logical reference chain rules; (2) MLK-LJP (Multi-Law Knowledge Enhanced Language Model for Legal Judgment Prediction), which employs a pure text legal knowledge graph for OCR text classification; (3) DocThinker, a general multimodal document understanding model that can be retrained as a document-level classifier.

Figure 6 shows the overall performance metrics of this paper and displays a clear lead over the first evaluation baseline across all performance metrics. The accuracy of 93.0% and the F1 score of 93.1% provide an improvement of more than 12 percentage points when compared to the MLK-LJP pure text classification model, confirming that graph network-based and reinforcement learning-based reasoning mechanisms need to incorporate visual layout and structured table information when trying to accurately identify the logical connections within legal documents. Specifically, these reasoning mechanisms can substantially reduce false negative results and cover complex logic scenarios that cannot easily be incorporated by rules alone. According to

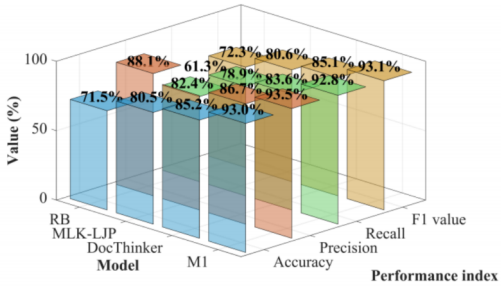

FIGURE 6. Overall performance comparison of end-to-end compliance verification tasks

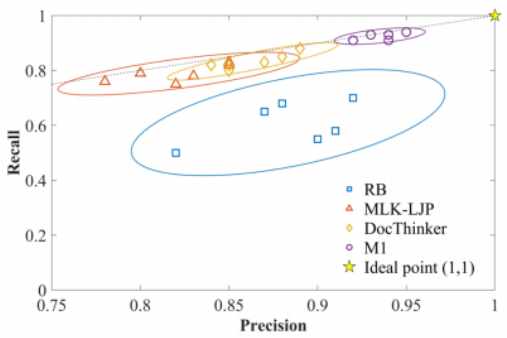

FIGURE 7. Performance distribution of different models on various compliance verification problems

the results from the current study and baseline comparison group, there is significant variability in the ability of all models to address a variety of compliance-related problems; specifically, the current study framework's recall rate is 31.5 percentage points higher than that of the baseline of rules. Figure 7 illustrates the performance of each model on six different types of compliance verification sub-tasks using both precision and recall as measures on the x-axis and y-axis respectively. Each of the six data points represents a single sub-task; the shape of the data points is used to identify each of the different models. The user-defined 95% confidence ellipse is a visual representation of both the concentration and stability of each model's overall performance.

The evaluation of various models demonstrates various performance distribution types, as illustrated in Figure 7. A concentration of rule baseline model (represented with a blue square) scatter point data indicates highprecision, low-recall (i.e., strict adherence to rules yet with an incomplete ability to provide full coverage) performance at the lower right of Figure 7. Both elliptical regions for the pure textual model MLK-LJP (represented by an upper red triangle) and general model DocThinker (represented as a yellow rhombus) overlap in physical location mid-graph exhibit relatively similar levels of recall and precision due illustrated position. The proposed model M1 (indicated with a purple circle), having the greatest concentration upwards towards the right in Figure 7, is most likely to produce high recall and high-precision outputs across the accuracy of all compliance related issues. This further supports the comprehensive advantages and overall robust capability of the proposed framework in performing complex logic verification evaluations across many compliance types.

In summation, multimodal data fusion and graph-based reasoning principles discussed in this paper advance the accuracy level of extracting tabular data from legal documents. The literature provides evidence that end-to-end compliance assessment through structured knowledge association, dynamic reasoning path optimization, and multimodality of data fusion have led to dual breakthroughs in relation to confidence, accuracy, and recall for compliance assessment evaluations. The model provides an effective technology solution for solving critical challenges related to the automated review process within legally mandated documentation systems.

D. ABLATION EXPERIMENTS

The intent behind this section is to conduct a systematic ablation experiment to determine whether each of the major components in the system framework, namely: multimodal fusion, GNN modeling, reinforcement learning optimization, and modality aware attention, is necessary to solving the compliance verification problem. The ablation experiment is carried out using the public dataset DocLayNet for simulating the compliance verification problem. The ablation experiment consists of five model variations: A) full framework; B) text only (no multimodal fusion); C) no GNN; D) fixed rule reasoning (no reinforcement learning optimization); E) no modality aware attention. Each of the five model variations is trained and tested on the same training and testing sets, and all five model variations are evaluated based on metrics pertaining to GritS used to measure table structure reconstruction, as well as performance metrics used for evaluating end-to-end compliance verification. The results from the ablation experiment are presented in Table 1.

TABLE 1. Ablation experiment results

Model variants	Table restoration (GriTS)	Compliance Verification		
		Accuracy (%)	Recall (%)	F1 value (%)
A	0.927	93.0	92.8	93.1
B	0.831	84.6	82.5	83.5
C	0.927	89.1	85.4	87.2
D	0.927	90.5	86.7	88.6
E	0.916	92.1	91.3	91.7

The complete model (A) outperforms all other models on both core metrics (GriTS = 0.927, F1 = 93.1%) as shown in Table 1. In particular, variant B suffers the greatest drop in performance with its compliance verification F1 score declining the greatest amount to 83.5%, demonstrating the essential nature of visual and layout information for understanding legal documents accurately, particularly when parsing their complex table structures. Variant C also suffers a drop in its compliance verification F1 score to 87.2% without impacting its table reconstruction accuracy illustrating the importance of explicitly modeling complex relationships between entities and clauses using graph networks to arrive at correct logical reasoning. Variant D has significant drops in both recall and F1 scores, indicating that the use of static rule paths is too restrictive, while the reinforcement learning-based dynamic optimization mechanism has significantly increased the system's recall capacity. Variant E also has a small but measurable drop in performance across all metrics confirming that this mechanism successfully enhances the refined fusion of heterogeneous feature sets and contributes significantly to the overall performance of the final model.

E. EFFICIENCY IMPROVEMENT AND PRACTICAL APPLICATION VALUE ASSESSMENT

In order to assess how much better efficiency has improved and to provide an overall value of the framework discussed in practical use cases, this section uses the DocLayNet publicly available dataset to create an efficiency test set, using a systematic comparison to four methods of conducting a review, as follows: 1) traditional manual review (Baseline), 2) rule-based approach (CPU Based), 3) method used in this paper (CPU inference only), 4) method used in this paper (GPU inference). There are 150 common legal documents in this test set with differing levels of complexity including tables and compliance verification types, ensuring there is a broad range of documents considered for evaluation. The evaluation of efficiency provides the results displayed in Table 2.

TABLE 2. Comparison of average processing efficiency per case for different review methods

Review Method	Average Processing Time (Minutes)	Time Standard Deviation (Minutes)	Speedup Factor Relative to Manual Review	Accuracy (%)	F1 Score (%)
Traditional Manual Review	240.0	45.6	1.0 (Baseline)	100.0*	100.0*
Rule-Based Baseline Method (CPU)	22.0	1.5	10.9	71.5	72.3
Proposed Method (CPU Inference Only)	8.2	0.3	29.3	93.0	92.8
Proposed Method (GPU Inference)	1.5	0.1	160.0	93.0	92.8

Note: * The accuracy and F1 score of traditional manual review are assumed to be at an ideal baseline of 100%.

Table 2 shows that the average processing time for traditional manual review is 240 minutes (4 hours), reflecting the inherently high intensity and time-consuming nature of legal document review. The rule-based baseline method reduces the average processing time to 22 minutes in a CPU environment, but its 71.5% accuracy still necessitates significant manual review in practical applications, limiting the efficiency improvement. The proposed method further reduces the average processing time to 8.2 minutes in a CPU environment and even reaches 1.5 minutes in a GPU-accelerated environment, representing speedups of 2.7 times and 14.7 times respectively compared to the rule-based baseline method, and 29.3 times and 160 times compared to manual review. This efficiency leap is mainly attributed to the framework's end-to-end automated process, particularly the dynamic optimization of the inference path by the reinforcement learning agent, avoiding numerous invalid pattern matching and search attempts in traditional rule-based systems.

The 160-fold speedup is calculated by comparing the complete end-to-end automated processing time against the total manual review time required for the same document, including document reading, table entry, cross-page information association, and clause verification. The 1.5 minute per document processing time on four A100 GPUs corresponds to a throughput of approximately 0.67 documents per minute. In industrial document processing contexts where throughput requirements vary widely, this performance level suits compliance verification scenarios that prioritize accuracy and interpretability over millisecond-level latency.

In order to demonstrate the improvement in efficiency provided by the framework described herein, a comparison of the time to process a document with the framework versus traditional manual review is provided in Table 3.

TABLE 3. Detailed time breakdown results for the same case

Processing Steps	Traditional Manual Review Time (s)	Proposed Framework Processing Time (ms)	Efficiency Improvement Factor (Manual/Framework)
Document Reading and Location	180	–	–
Table Information Extraction and Entry	300	–	–
OCR and Layout Analysis	–	380	–
Cross-Page Information Association	120	–	–
Multi-Modal Table Parsing	–	520	–
Amount Calculation and Verification	180	–	–
Graph Network Construction	–	150	–
Clause Search and Comparison	120	–	–
Reinforcement Learning Inference	–	150	–
Total	900	1200	750

Table 3 shows that manual review requires cross-page information retrieval, manual calculation, and clause verification, taking approximately 15 minutes (900 seconds) in total, with each step carrying the risk of human error. This framework, through end-to-end automation, reduces the total time to 1.2 seconds (1200 milliseconds). The reinforcement learning inference path optimization reduces the verification step time to only 1/750th of the manual time, achieving an order-of-magnitude improvement in efficiency.

IV. CONCLUSIONS

The multimodal legal document analysis framework developed in this study significantly improves the accuracy of structured information extraction from large and complex tables by integrating visual, textual,and layout data (particularly for nested or merged cells). The approach combines the GNN model of the entity-clause relationship with reinforcement learning optimization of the solving path, resulting in both very high overall accuracy and compliance verification (coverage) during automated processing. Although the GriTS metric effectively captures overall structural fidelity in table reconstruction, it does not differentiate between routine cells and those containing legally critical information such as limit values or clause numbers. Future work will introduce a Critical Field Error Rate (CFER) metric that specifically measures extraction errors in such high-stakes fields, thereby providing a more nuanced assessment of framework reliability in compliance-sensitive scenarios. Through rigorous testing, the framework has been proven to significantly enhance processing performance while maintaining high interpretability. This ultimately yields an efficient and intelligent end-to-end solution for smart legal document review , demonstrating broad practical value.

AUTHOR CONTRIBUTIONS

WeiZhong Tang designed, collected and analyzed the data, and drafted the manuscript. WeiZhong Tang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. WeiZhong Tang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by,
National Key Research and Development Program Project (2024YFC3306503).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] S. Joseph, “Advanced digital image processing technique based optical character recognition of scanned document,” Journal of Innovative Image Processing, vol. 4, no. 3, pp. 195-205, 2022, doi: 10.36548/jiip.2022.3.007.

[2] S. Mukherjee, H. Tyagi, P. Tyagi, et al., “OCR using python and its application,” Journal of Computer Science, vol. 12, no. 3, pp. 45-58, 2023, doi: 10.17762/jaz.v44iS-3.1062.

[3] A. Chawla, A. Gupta, and S. Shushrutha K, “Intelligent information retrieval: techniques for character recognition and structured data extraction,” J. Emerg. Technol. Innov. Res.(JETIR), vol. 9, no. 7, pp. 452-459, 2022.

[4] R. Samantapudi R K, “Table Extraction from Financial and Transactional Documents,” International journal of IoT, vol. 5, no. 01, pp. 95-125, 2025, doi: 10.55640/ijiot-05-01-06.

[5] K. Pappula K and P. Rusum G, “Multi-Modal AI for Structured Data Extraction from Documents,” International Journal of Emerging Research in Engineering and Technology, vol. 4, no. 3, pp. 75-86, 2023, doi: 10.63282/3050-922X.IJERET-V4I3P109.

[6] K. Pappula K, “Transformer-Based Classification of Financial Documents in Hybrid Workflows,” International Journal of Multidisciplinary on Science and Management, vol. 1, no. 3, pp. 48-61, 2024.

[7] Z. Zhang, K. Chen, R. Wang, et al., “Universal multimodal representation for language understanding,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 7, pp. 9169-9185, 2023, doi: 10.1109/TPAMI.2023.3234170.

[8] B. Oral and G. Eryiğit, “Fusion of visual representations for multimodal information extraction from unstructured transactional documents,” International Journal on Document Analysis and Recognition (IJDAR), vol. 25, no. 3, pp. 187-205, 2022, doi: 10.1007/s10032-022-00399-3.

[9] L. Polo, “The role of AI and OCR-based label verification systems in enhancing food traceability and supply chain transparency,” International Journal for Multidisciplinary Research, vol. 7, no. 2, pp. 1-18, 2025, doi: 10.36948/ ijfmr.2025.v07i02.41577.

[10] S. Dutta, S. Adhikary, and D. Dwivedi A, “Visformers-Combining vision and transformers for enhanced complex document classification,” Machine Learning and Knowledge Extraction, vol. 6, no. 1, pp. 448-463, 2024, doi: 10.3390/make6010023.

[11] Q. Zhang, "Enhanced Feature Fusion and Transfer Learning for Multi-Format Government Document Classification," Journal of Science, Innovation & Social Impact, vol. 1, no. 1, pp. 427-441, 2025.
[12] S. Maldini A, J. Saputra W S, and A. Prasetya D, "Multimodal Detection of Covert Online Gambling Advertisements Using Faster R-CNN and Tr-OCR," bit-Tech, vol. 8, no. 1, pp. 953-963, 2025, doi: 10.32877/bt.v8i1.2769.
[13] B. Bulatov K, V. Bezmaternykh P, P. Nikolaev D, et al., "Towards a unified framework for identity documents analysis and recognition," Kompyuternaya optika, vol. 46, no. 3, pp. 436-454, 2022, doi: 10.18287/2412-6179-CO-1024.
[14] S. Mandvikar, "Augmenting intelligent document processing (IDP) workflows with contemporary large language models (LLMs)," International Journal of Computer Trends and Technology, vol. 71, no. 10, pp. 80-91, 2023, doi: 10.14445/22312803/IJCTT-V71I10P110.
[15] T. Odofin, A. Abayomi A, E. Ogbuefi, et al., "Strategic integration of LangChain, Hugging Face Transformers, and OpenAI for document intelligence systems," International Journal of Scientific Research in Science Engineering and Technology, vol. 11, no. 4, pp. 413-426, 2024, doi: 10.32628/IJSRSET25121177.
[16] V. Naik, P. Patel, and R. Kannan, "Legal entity extraction: An experimental study of NER approach for legal documents," International Journal of Advanced Computer Science and Applications, vol. 14, no. 3, pp. 775-783, 2023, doi: 10.14569/ IJACSA.2023.0140389.
[17] B. Sinha, S. Saxena, A. Vashishtha, et al., "Legal Ease: An End-to-End Automated Legal Document Processing System Using LLMs and OCR," International Journal of Software Computing and Testing, vol. 11, no. 1, pp. 29-32p, 2025.
[18] J. Chen, M. Wang, and T. Sun, "Intelligent Tax Systems and the Role of Natural Language Processing in Regulatory Interpretation," American Journal of Machine Learning, vol. 6, no. 4, pp. 74-94, 2025, doi: 10.71465/ajml3452.
[19] T. Zhang, "A Knowledge Graph-Enhanced Multimodal AI Framework for Intelligent Tax Data Integration and Compliance Enhancement," Frontiers in Business and Finance, vol. 2, no. 02, pp. 247-261, 2025, doi: 10.71465/fbf384.
[20] V. Bezditnyi, "Use of artificial intelligence for tax planning optimization and regulatory compliance," Research Corridor Journal of Engineering Science, vol. 1, no. 1, pp. 103-142, 2024, doi: 10.66320/8qe5f566.
[21] A. Parakala and R. Pothula, "AI+ Document Understanding in UiPath: Solving Real Government Problems," International Journal of Artificial Intelligence, Data Science, and Machine Learning, vol. 3, no. 3, pp. 111-122, 2022, doi: 10.63282/3050-9262.IJAIDSML-V3I3P112.
[22] R. Pingili, "AI-driven intelligent document processing for banking and finance," International Journal of Management & Entrepreneurship Research, vol. 7, no. 2, pp. 98-109, 2025, doi: 10.51594/ijmer.v7i2.1802.
[23] S. Talakola, "Transforming BOL Images into Structured Data Using AI," International Journal of Artificial Intelligence, Data Science, and Machine Learning, vol. 6, no. 1, pp. 105-114, 2025, doi: 10.63282/3050-9262.IJAIDSML-V6I1P112.
[24] A. Chansarkar, "Enhancing Unstructured Document Text Extraction with LLMs in Cloud-Native Enterprise Solutions," Journal Of Engineering And Computer Sciences, vol. 4, no. 9, pp. 251-257, 2025.
[25] P. Sharma, "Advancements in OCR: a deep learning algorithm for enhanced text recognition," International Journal of Inventive Engineering and Sciences, vol. 10, no. 8, pp. 1-7, 2023, doi: 10.35940/ijies.F4263.0810823.
[26] K. Soni V, V. Shukla, R. Tandan S, et al., "Performance Evaluation of Efficient and Accurate Text Detection and Recognition in Natural Scenes Images Using EAST and OCR Fusion," International Journal of Advanced Computer Science & Applications, vol. 16, no. 1, pp. 445-453, 2025, doi: 10.14569/IJACSA.2025.0160144.
[27] H. Gbada, K. Kalti, and A. Mahjoub M, "Deep learning approaches for information extraction from visually rich documents: datasets, challenges and methods," International Journal on Document Analysis and Recognition (IJDAR), vol. 28, no. 1, pp. 121-142, 2025, doi: 10.1007/s10032-024-00493-8.
[28] F. Wang, S. Tian, L. Yu, et al., "TEDT: transformer-based encoding-decoding translation network for multimodal sentiment analysis," Cognitive Computation, vol. 15, no. 1, pp. 289-303, 2023, doi: 10.1007/s12559-022-10073-9.
[29] L. Sun, Z. Lian, B. Liu, et al., "Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis," IEEE Transactions on Affective Computing, vol. 15, no. 1, pp. 309-325, 2023, doi: 10.1109/ TAFFC.2023.3274829.
[30] P. Xu, X. Zhu, and A. Clifton D, "Multimodal learning with transformers: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 10, pp. 12113-12132, 2023, doi: 10.1109/TPAMI.2023.3275156.
[31] M. Binte Rashid, S. Rahaman M, and P. Rivas, "Navigating the multimodal landscape: A review on integration of text and image data in machine learning architectures," Machine Learning and Knowledge Extraction, vol. 6, no. 3, pp. 1545-1563, 2024, doi: 10.3390/make6030074.
[32] F. Shi, R. Gao, W. Huang, et al., "Dynamic mdetr: A dynamic multimodal transformer decoder for visual grounding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 2, pp. 1181-1198, 2023, doi: 10.1109/ TPAMI.2023.3328185.
[33] J. Li, Y. Wang, and D. Zhao, "Layer-wise enhanced transformer with multi-modal fusion for image caption," Multimedia Systems, vol. 29, no. 3, pp. 1043-1056, 2023, doi: 10.1007/s00530-022-01036-z.
[34] H. Zhang, B. Wu, X. Yuan, et al., "Trustworthy graph neural networks: Aspects, methods, and trends," Proceedings of the IEEE, vol. 112, no. 2, pp. 97-139, 2024, doi: 10.1109/JPROC.2024.3369017.
[35] B. Abimbola, E. de La Cal Marin, and Q. Tan, "Enhancing legal sentiment analysis: A convolutional neural network-long short-term memory document-level model," Machine Learning and Knowledge Extraction, vol. 6, no. 2, pp. 877-897, 2024, doi: 10.3390/make6020041.
[36] A. Yusuf A, F. Chong, and M. Xianling, "An analysis of graph convolutional networks and recent datasets for visual question answering," Artificial Intelligence Review, vol. 55, no. 8, pp. 6277-6300, 2022, doi: 10.1007/s10462-022-10151-2.
[37] X. Liu, C. Miao, G. Fiumara, et al., "Information propagation prediction based on spatial-temporal attention and heterogeneous graph convolutional networks," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 945-958, 2023, doi: 10.1109/TCSS.2023.3244573.