Reading the Signs: A Graph-Based System for Multimodal Information Retrieval on Vietnamese Traffic Law

Hieu Minh Huynh [1,3], **An Nguyen Tran Khuong**[1,3], **Dai Phan Trong**[2,3], **Tin Van Huynh**[1,3]
[1]University of Information Technology, Ho Chi Minh City, Vietnam
[2]University of Science, Ho Chi Minh City, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam
{22520434,22520026}@gm.uit.edu.vn, 23120026@student.hcmus.edu.vn, tinhv@uit.edu.vn

Abstract

The interpretation of traffic signs in accordance with legal statutes is a challenging multimodal reasoning task, as it requires integrating visual recognition with statutory provisions. This paper presents our system for the VLSP 2025 shared task on Multimodal Legal Question Answering on Traffic Sign Rules (MLQA-TSR). We propose a graph-based retrieval framework that explicitly models relationships among legal articles, traffic sign images, and tabular data. Grounding DINO is used for zero-shot traffic sign detection, and SigLIP encodes visual features for alignment with a heterogeneous knowledge graph. To refine relevance, we incorporate the jinaai/jina-reranker-m0 model as a multimodal reranker for traffic sign selection. These signs serve as entry points for cosine similarity search and structured traversal with a dynamic top-k strategy to retrieve legal context. For question answering, we design a three-stage pipeline that integrates image processing, concise legal context extraction, and final reasoning. The reasoning stage ensembles Qwen2.5-VL-7B-Instruct and InternVL3-8B models with chain-of-thought and self-reflection prompting to improve accuracy and interpretability. Experiments on the official benchmark confirm the effectiveness of our system for multimodal legal QA, highlighting the advantages of structured graph-based retrieval combined with multimodal reranking and reasoning.

1 Introduction

The rapid increase of traffic in urban areas, particularly in Vietnam, has led to a complex system of traffic laws and regulations. For drivers, interpreting traffic signs in real-world scenes and understanding their legal implications is a significant challenge. This paper addresses the VLSP 2025 challenge on Multimodal Legal QA on Traffic Sign Rules (VLSP 2025 MLQA-TSR). The challenge is divided into two sequential subtasks: first, retrieving relevant legal articles (Task 1), and second, answering a specific question using the provided context (Task 2). Per the shared task design, the gold labels from Task 1 are provided as input for Task 2, allowing for isolated evaluation of the VQA component.

Our primary contribution is a novel, graph-based multimodal system. For the retrieval task, we construct a heterogeneous knowledge graph that explicitly connects law articles, traffic sign images, and tabular data to robustly model the legal domain. Our query processing pipeline uses a multimodal reranker to identify salient visual evidence, followed by a similarity search and structured graph traversal to pinpoint relevant legal context. For the question answering task, the highly accurate articles identified by our retrieval process serve as a grounded knowledge base for a VQA module that employs chain-of-thought reasoning to derive the final answer.

2 Related Work

Multimodal Information Retrieval. Recent advances in multimodal QA highlight the benefits of retrieval-augmented systems that combine textual and visual evidence. Approaches such as MuRAG (Chen et al., 2022) and MuRAR (Zhu et al., 2025b) retrieve relevant image–text pairs to improve grounding, while M3DocRAG (Cho et al., 2024) extends this paradigm to multi-page documents with figures and tables. These works demonstrate that multimodal retrieval substantially improves QA performance over text-only baselines.

In the legal domain, retrieval-augmented generation has been shown to reduce hallucination and improve statutory interpretation and case law reasoning (Kabir et al., 2025). Existing efforts largely focus on textual pipelines, leveraging case-based retrieval and legal knowledge graphs to organize statutes and precedents. However, these systems remain predominantly text-centric, overlooking the role of visual evidence that is critical in domains such as traffic law, where legal interpretation depends on both statutory provisions and physical signs. This gap motivates multimodal approaches that explicitly link visual cues (e.g., traffic signs) to legal rules, ensuring that answers are grounded not only in textual statutes but also in the visual context of the question.

Graph-based approaches further enrich retrieval by exploiting structured relations. Symbolic traversal over knowledge graphs enables explicit multi-hop reasoning, while neural methods such as GNN-Ret (Li et al., 2024) and CaseGNN (Tang et al., 2023) leverage graph neural networks to capture inter-document dependencies. More recently, multimodal knowledge graphs (e.g., mKG-RAG (Yuan et al., 2025)) unify text and images within retrieval-augmented pipelines, though applications in the legal setting remain largely unexplored.

Our work builds on these directions by introducing a graph-based multimodal retrieval system tailored for traffic law QA. Unlike prior legal QA approaches that rely mainly on textual retrieval, we explicitly connect traffic sign images, tabular data, and legal articles in a heterogeneous knowledge graph, enabling structured traversal and grounded multimodal reasoning.

Visual Question Answering. Complementing retrieval and graph methods, recent progress in open-source vision–language backbones and prompting strategies has materially advanced VQA performance. In particular, Qwen2.5-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025a) demonstrate strong multimodal understanding and provide practical, high-quality bases for downstream reasoning. At the prompting level, Chain-of-Thought (CoT) prompting has been shown to elicit step-wise reasoning in large models (Wei et al., 2023), and subsequent work indicates that concise or self-reflective reasoning traces can further improve problem-solving fidelity and reduce spurious outputs. Motivated by these findings, our VQA pipeline combines retrieval-grounded context with CoT and self-reflection style prompting to improve factual grounding and reduce hallucination in vision–language reasoning.

3 Method

Our overall system is composed of two sequential pipelines designed to address the two subtasks. Following the shared task's structure, the VQA module is developed and evaluated using the gold-standard retrieved articles provided by the organizers after the completion of Task 1.

3.1 Subtask 1: Multimodal Information Retrieval

Our proposed system for the information retrieval subtask is a comprehensive pipeline structured into three main stages: Data Processing and Feature Extraction (3.1.1), Heterogeneous Graph Construction (3.1.2), and Multi-Modal Query Processing and Retrieval (3.1.3), as illustrated in Figure 1

3.1.1 Data Processing and Feature Extraction

The initial stage focuses on processing the raw visual data from both the legal document corpus and the user queries to create a standardized set of features. We employ Grounding DINO to automatically identify and crop traffic signs from all images. To enable semantic comparison, all cropped signs are then encoded into high-dimensional vector representations using the SigLIP model.

3.1.2 Heterogeneous Graph Construction

To model the rich, structured information and explicit relationships within the legal corpus, we construct a heterogeneous knowledge graph. This graph is foundational to our retrieval strategy, providing a structured map of the legal domain. The construction process is entirely rule-based, ensuring deterministic and accurate links between different types of information. The graph comprises three distinct types of nodes and three types of edges connecting them.

- **TextNode**: Represents an individual law article, containing its title and processed textual content.
- **ImageNode**: Represents a single, specific traffic sign that has been cropped from the images in the law database.
- **TableNode**: Represents a structured table extracted from within the text of a law article.

The connections between nodes are not learned but are created by parsing the explicit structure and text of the legal documents. This deterministic approach guarantees that the graph accurately reflects the citations and references in the source material.

- **Text-Image** and **Text-Table Edges**: These edges connect an article to the visual or tabular data it explicitly contains. They are created

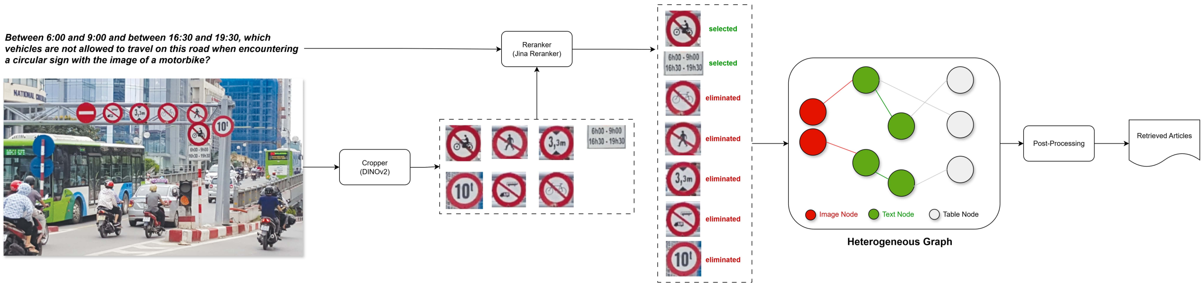
Between 6:00 and 9:00 and between 16:30 and 19:30, which vehicles are not allowed to travel on this road when encountering a circular sign with the image of a motorbike?
Reranker (Jina Reranker)
Cropper (DINOv2)
6h00 - 9h00 16h30 - 19h30
selected
selected
eliminated
eliminated
eliminated
eliminated
eliminated
Image Node
Text Node
Table Node
Heterogeneous Graph
Post-Processing
Retrieved Articles

Figure 1: Overview of our retrieval system, which integrates data processing, heterogeneous graph construction, and multi-modal query processing for effective information retrieval.

by parsing markup tags within the article's text. An edge is formed between a TextNode and an ImageNode if the text contains an <IMAGE:...> tag referencing that specific sign image. Similarly, a Text-Table edge is created when a <TABLE:...> tag is found. This creates an undirect, many-to-many link between articles and their embedded visual and tabular elements.

- **Text-Text Edges**: To capture the intricate web of inter-article citations, we establish undirected edges between TextNodes using syntactic pattern matching. We apply a set of regular expressions to the text of each article to find explicit references to other articles, clauses, or appendices. The primary patterns include Điều X (Article X), Khoản Y (Clause Y), and Phụ lục [A-Z] (Appendix [A-Z]). When a pattern is matched—for instance, if the text of Article 15 contains a reference to "Điều 21"—an undirected edge is created from the TextNode for Article 15 to the TextNode for Article 21. This rule-based system transforms the flat legal corpus into a structured knowledge graph where legal citations are represented as traversable paths.

3.1.3 Multi-Modal Query Processing and Retrieval

This final stage processes a user's query (image and text) to retrieve the relevant legal articles. First, we use the jinaai/jina-reranker-m0 model(AI, 2024) to identify which traffic signs in the scene image are relevant to the question. The relevant signs are then used to perform a cosine similarity search against the law database's sign embeddings to find entry points (ImageNodes) into our graph (see Figure. 2). Starting from these nodes, we perform a Breadth-First Search (BFS) traversal to find connected TextNodes. Finally, to ensure the number of retrieved articles is proportional to the complexity of the query, we apply a dynamic top-k selection strategy. We observed that legal articles in the corpus often appear in pairs: one article describing a general rule, and another providing specific details. To capture this structure, the number of results to return, k, is determined by the number of relevant signs $(n\_cropped)$ identified by the reranker, using the formula $k = 2 \times n\_cropped$. This heuristic is designed to retrieve the likely pair of articles for each relevant sign, with the goal of improving the F2 score by balancing precision and recall.

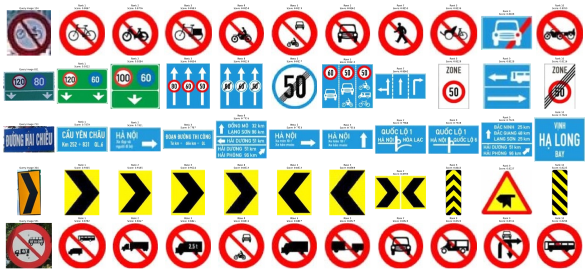

Figure 2: Visual similarity search results. The leftmost column shows query images cropped from scenes. The subsequent columns display the top images retrieved from the law database, which are used to map the query to a corresponding ImageNode

3.2 Subtask 2: Visual Question Answering

Our Visual Question Answering (VQA) module is implemented as a three-stage pipeline (Figure 3) that integrates questions, visual evidence, legal text, and structured reasoning to produce the final answer. The stages are: (i) *Image Conditioning* (3.2.1) , which prepares visual inputs for clarity; (ii) *Law Context Extraction* (3.2.2), which distills a concise legal context based on the question and retrieved articles; and (iii) *Reasoning and Answer Selection* (3.2.3), which generates the final answer using vision-language models.

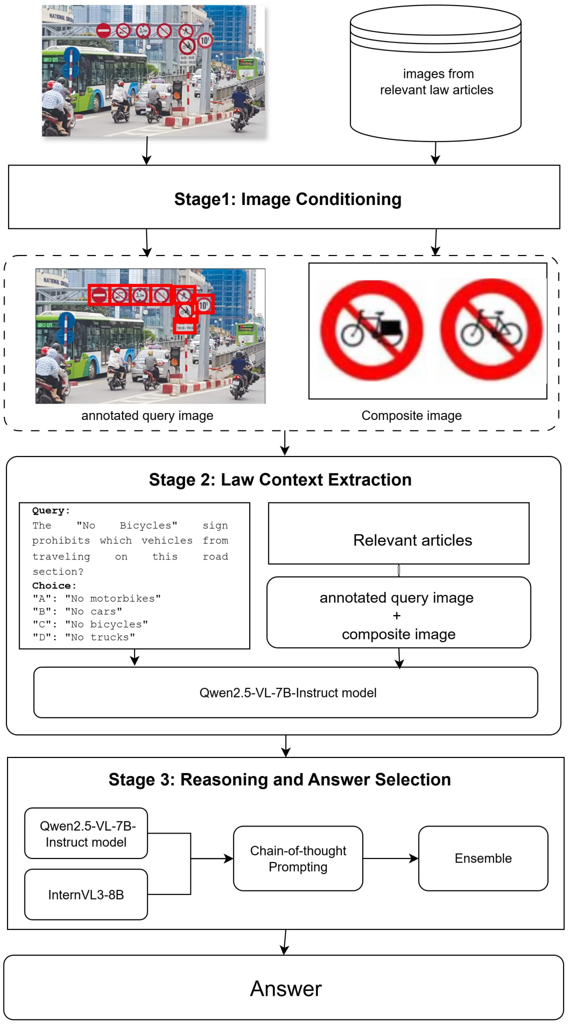
images from relevant law articles
Stage1: Image Conditioning
annotated query image
Composite image
Stage 2: Law Context Extraction
Query:
The "No Bicycles" sign prohibits which vehicles from traveling on this road section?
Choice:
"A": "No motorbikes"
"B": "No cars"
"C": "No bicycles"
"D": "No trucks"
Relevant articles
annotated query image + composite image
Qwen2.5-VL-7B-Instruct model
Stage 3: Reasoning and Answer Selection
Qwen2.5-VL-7B-Instruct model
InternVL3-8B
Chain-of-thought Prompting
Ensemble
Answer

Figure 3: Overview of the Visual Question Answering (VQA) pipeline. The system combines annotated query images, retrieved law articles, and structured reasoning prompts.

3.2.1 Image Conditioning

To prepare the visual evidence, we apply two preprocessing steps. First, for each retrieved law article, all associated images are merged into a single composite image. Second, the user's query image is annotated using Grounding DINO, which detects traffic signs and overlays bounding boxes to highlight relevant regions. These annotated images provide clear attention anchors for downstream models.

3.2.2 Law Context Extraction

This stage generates a focused legal context conditioned on the visual evidence. We employ QwenVL, which receives the question, annotated query image, retrieved law text, and composite law image. The model is instructed to extract only the details relevant to the detected traffic signs. To avoid premature bias, answer options are withheld from the model at this stage.

3.2.3 Reasoning and Answer Selection

The final stage derives the answer from the curated inputs. We evaluate two small vision language models, QwenVL2.5-7B and InternVL3-8B. Each model is provided with the question, the candidate answer choices, an annotated query image, and the extracted legal context. To promote transparent deliberation, we apply chain-of-thought and self-reflection prompting. Finally, we aggregate the predictions of both models through an ensemble strategy to enhance robustness and reduce variance in answer selection.

We experiment with two prompt variants, P_{detail} and P_{simple}. Figure 4 presents the prompt template for P_{detail}, which enforces a structured three step reasoning schema.

Role: Vietnamese law expert
General Instruction:
1) Analyse the question;
2) Explain reasoning behind the choice;
3) Provide final answer.
Additional Information:
– Signs present in the image
– Exception: priority is always granted to ambulances and police cars
Output Format:

<reasoning> ... </reasoning>
<answer> ... </answer>

Input: – Question – Answer choices

Figure 4: Prompt template for P_{detail}. P_{simple} differs only in its general instruction, which directs the model to "think step by step".

$P_{\textbf{detail}}$ The output format explicitly guides the model through three distinct steps: Analyse the question - Analyse the image - Reasoning based on the provided choices

$P_{\textbf{simple}}$ The output format only requires the model to *"think step by step"* before producing the final answer.

4 Experiment Result

4.1 Dataset

The VLSP 2025 MLQA-TSR dataset consists of two main parts: a legal database and multimodal QA pairs.

Law database. The legal database covers two official Vietnamese traffic regulations: *QCVN 41:2024/BGTVT* with 310 articles and *Law on Road Traffic Order 36/2024/QH15* with 89 articles, totaling 399 articles. Many articles include structured content such as 762 images and 212 tables, with references heavily skewed toward *QCVN 41:2024/BGTVT* (99.3%).

QA pairs. The dataset provides 530 training questions and 196 test questions (50 public, 146 private). Questions appear in two formats: multiple-choice (70.9%) and yes/no (29.1%). The answer distribution is moderately balanced, as summarized in Table 1. On average, each question contains 18.8 words. In total, 130 unique legal articles and 419 distinct traffic signs are explicitly linked to the corpus, supporting grounded multimodal reasoning.

We do not use this data to train any model but only as a validation set.

Answer label	Proportion (%)
A	21.7
B	19.4
C	15.8
D	14.0
Đúng (Yes)	14.2
Sai (No)	14.9

Table 1: Answer label distribution in the VLSP 2025 MLQA-TSR training set.

4.2 Evaluation Metrics

4.2.1 Evaluation Metrics – Retrieval Task

Task 1 is evaluated using precision, recall, and the F2 score, which emphasizes recall to better capture the relevance of retrieved legal articles. For each question i, the metrics are computed as:

$$\text{precision}_i = \frac{\#\text{correctly retrieved articles}}{\#\text{retrieved articles}},$$

$$\text{recall}_i = \frac{\#\text{correctly retrieved articles}}{\#\text{relevant articles}}.$$

The F2 score for question i is then defined as:

$$\text{F2}_i = \frac{5 \times \text{precision}_i \times \text{recall}_i}{4 \times \text{precision}_i + \text{recall}_i}.$$

The final evaluation metric is the macro-averaged F2 score across all N questions:

$$\text{F2}_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \text{F2}_i.$$

This macro-averaging ensures that each question contributes equally to the overall score, preventing bias toward queries with more associated legal references.

4.2.2 Evaluation Metrics – VQA Task

Task 2 is evaluated using accuracy, a straightforward and interpretable metric that quantifies the proportion of correctly answered questions. It is defined as:

$$\text{accuracy} = \frac{\#\text{correctly answered questions}}{\#\text{total questions}}.$$

Accuracy directly captures the end-user utility of the system, as it reflects the frequency with which the model provides the correct legal judgment. This is particularly relevant for legal QA, where interpretability and correctness of the final answer are critical.

4.3 Results

4.3.1 Subtask 1. Multimodal Information Retrieval

Our system ranked **2nd** on the official leaderboard with an F2 score of 0.611, slightly behind the top-ranked team (0.646), as shown in Table 2.

Table 2: VLSP 2025 MLQA-TSR leaderboard results for Subtask 1 (Legal Retrieval). Scores are reported in terms of F2.

Team name	F2 Score	Rank
life_is_tough	0.646	1
Ours	0.611	2
tieen	0.599	3

To better understand the effectiveness of our methods, we further conducted experiments on the training set. We compared different graph traversal strategies (Depth-First Search vs. Breadth-First Search) and analyzed the impact of our dynamic top-k selection strategy against a fixed-k baseline. The results, measured in Precision (P), Recall (R), and F2-Score, are presented in Table 3.

Table 3: Evaluation Results on Training Set

Experiment	P	R	F2
Baseline (BGE-m3)	-	-	0.12
DFS, K=3 (best)	0.4132	0.5657	0.5136
DFS, w/ Dynamic-K	**0.5472**	0.5081	0.5061
BFS, K=3 (best)	0.4145	0.5676	0.5153
BFS, w/ Dynamic-K	0.4865	**0.6176**	**0.5503**

The results show that all graph-based methods significantly outperform the baseline. While both Breadth-First Search (BFS) and Depth-First Search (DFS) are effective, BFS consistently yields slightly better results. The most significant improvement comes from our **dynamic top-k selection strategy**, particularly when paired with BFS, which achieved the highest recall (0.6176) and the best overall F2-score (0.5503). Because the F2-score prioritizes recall, the BFS w/ Dynamic-K configuration was selected for our final submission.

Finally, we evaluated this best-performing configuration on the official test set. As shown in Table 4, BFS with dynamic top-k achieved strong performance, with recall reaching 0.6501 and F2-score 0.6114, which aligns with our official leaderboard result in Table 2. This confirms the robustness of our retrieval strategy across both validation and test sets.

Table 4: Evaluation results of BFS with dynamic top-K on the test set. The row ALL corresponds to using the dynamic top-K strategy across all queries.

K	Precision	Recall	F2-Score
1	0.8151	0.3344	0.3759
2	0.6986	0.5699	0.5829
3	0.5091	0.6124	0.5763
4	0.3955	0.6271	0.5473
5	0.3192	0.6289	0.5116
6	0.2728	0.6409	0.4887
7	0.2348	0.6426	0.4610
8	0.2082	0.6500	0.4386
9	0.1865	0.6491	0.4178
10	0.1685	0.6501	0.3980
15	0.1123	0.6501	0.3193
20	0.0842	0.6501	0.2671
30	0.0562	0.6501	0.2018
50	0.0337	0.6501	0.1359
100	0.0168	0.6501	0.0750
500	0.0034	0.6501	0.0164
ALL	0.6130	**0.6501**	**0.6114**

4.3.2 Subtask 2. Visual Question Answering

For the VQA task, our system achieved an accuracy of 0.623, placing us **8th** overall on the leaderboard (Table 5).

We also performed ablation experiments to quantify the contribution of different components. We evaluate the pipeline under the following configurations: (i) using only the reasoning stage, (ii) reasoning with annotated query images, (iii) reasoning with law context extraction, and (iv) the full three-stage pipeline.

Table 5: VLSP 2025 MLQA-TSR leaderboard results for Subtask 2 (Visual Question Answering). Scores are reported in terms of Accuracy.

Team name	Accuracy	Rank
dinhanhx	0.863	1
brownyeyez	0.836	2
tieen	0.781	3
Ours	0.623	8

Table 6 reports accuracy across different settings. Incorporating law context extraction consistently improves performance, as it directs the SVLMs to focus on the relevant legal rules. Annotating the query image provides an additional but smaller improvement. The ensemble of QwenVL and InternVL yields the best overall accuracy.

Table 6: Task 2 results on Visual Question Answering. Accuracy (Acc) is reported for different pipeline variants. The best score is in **bold**.

Base	Prompt	Annotation	Law Ext.	Acc
QwenVL	P_{simple}	No	No	0.613
QwenVL	P_{detail}	No	No	0.615
QwenVL	P_{detail}	Yes	No	0.619
QwenVL	P_{detail}	No	Yes	0.624
QwenVL	P_{detail}	Yes	Yes	0.626
InternVL	P_{detail}	Yes	Yes	0.625
QwenVL	P_{simple}	Yes	Yes	0.621
InternVL	P_{simple}	Yes	Yes	0.628
Ensemble	-	-	-	**0.632**

5 Discussion

5.1 Discussion

Our experimental analysis underscores two central findings: (i) Breadth-First Search (BFS) consistently outperforms Depth-First Search (DFS) and (ii) the proposed dynamic top-k strategy provides a principled improvement over a fixed-k baseline

Superiority of BFS over DFS. An examination of the VLSP dataset reveals that each question is paired with a single query image, which typically contains only a limited number of traffic signs relevant to the legal reasoning process. Each sign corresponds to one appendix in the legal corpus, and each appendix is uniquely linked to a single governing article. Within this structured hierarchy, BFS is particularly well suited. Starting from an ImageNode, BFS rapidly identifies the corresponding appendix (TextNode) and subsequently the governing article, thereby maintaining focus on the most relevant legal provisions. In contrast, DFS tends to traverse deeply into articles beyond the immediately relevant ones, often retrieving extraneous content that introduces noise. This structural alignment explains why BFS achieves superior recall and F2 performance across our experiments.

Effectiveness of dynamic top-k selection. A second empirical observation is that the regulation of a given traffic sign typically involves two complementary provisions: a general article and a corresponding appendix. A fixed-k retrieval strategy is inherently limited in accommodating this variability—setting k too low risks omitting pertinent articles (lowering recall), while setting k too high introduces irrelevant content (lowering precision). The dynamic top-k approach mitigates this limitation by adjusting the number of retrieved articles to the number of detected signs in the query image. Concretely, for n_{cropped} detected signs, we set $k = 2 \times n_{\text{cropped}}$, which approximates the true number of governing provisions. This adaptive mechanism strikes a more effective balance between precision and recall, as reflected in the observed gains in both recall and F2 score. These results demonstrate that dynamic top-k is not merely heuristic but a principled adaptation to the structural characteristics of traffic law data.

Although our study is grounded in Vietnamese traffic law, the observed advantages of BFS over DFS and the dynamic top-k strategy arise from structural characteristics of legal corpora such as hierarchical referencing and interdependent provisions, which are prevalent across jurisdictions and indicate the broader applicability of our approach to multimodal legal QA in diverse regulatory contexts.

5.2 Error Analysis

Failed case of Retrieval task Our analysis identified two primary sources of retrieval failures: the multimodal reranker and the visual similarity search model. The reranker can fail due to a lack of global context or confusion between similar signs, leading to two outcomes (see Table. 7). Over-ranking, as seen in query private_test_1, adds too many irrelevant signs, causing our dynamic top-k strategy to retrieve excess articles and harm precision. Conversely, under-ranking, as in query private_test_2, misses relevant signs, leading to the omission of correct articles and reducing recall.

Table 7: Examples of reranker failures. Incorrectly retrieved or missed articles are marked with an asterisk (*).

Query ID	Predicted Articles	Ground Truth
private_test_1 (Over-ranking)	QCVN 41, Art. 22 QCVN 41, Art. B.4 QCVN 41, Art. 36* QCVN 41, Art. E.14* QCVN 41, Art. E.48* QCVN 41, Art. E.17* QCVN 41, Art. 25* QCVN 41, Art. 73*	QCVN 41, Art. 22 QCVN 41, Art. B.4
private_test_2 (Under-ranking)	QCVN 41, Art. 41 QCVN 41, Art. F.10	QCVN 41, Art. 22* QCVN 41, Art. B.3* QCVN 41, Art. 41 QCVN 41, Art. F.10

Furthermore, the visual similarity search sometimes fails by prioritizing general visual features (shape, color) over the specific symbols that define a sign's legal meaning. For instance, a query for a "speed bump" sign incorrectly retrieves other rectangle information signs like "market", while a query for a supplementary sign retrieves other blue rectangular signs with different legal interpretations (Figure. 5). This highlights the model's limitation in fine-grained semantic differentiation, where visual similarity does not guarantee legal equivalence.

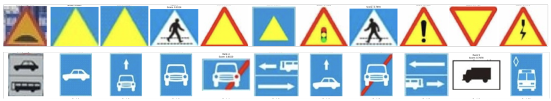

Figure 5: Example of a similarity search failure.

Failed Case of VQA task Our analysis identified three distinct patterns of error, even in seemingly simple tasks. First, the models struggle with spatial reasoning: when questions require distinguishing the relative position of traffic signs (e.g., left vs. right), the predictions often reveal confusion, despite the visual cues being explicit. Second, the models underperform on basic symbolic reasoning. For instance, when asked to determine parking permission on a specific calendar day, the models correctly identify the day but fail to map it to an odd/even category, leading to incorrect conclusions (see Table 8). Third, we observe reasoning–answer mismatches: although the chain-of-thought (CoT) rationale is correct, the final selected choice is inconsistent with the reasoning, even when self-reflection or randomized answer formatting is applied. These observations suggest that, rather than incrementally refining decisions through CoT prompting, the models often commit to an answer prematurely during question parsing. This highlights an important limitation of current small VLMs.

Table 8: Example of a Symbolic Reasoning Failure in the VQA Module. The model correctly identifies the rule for sign P.131c but incorrectly classifies the date ’17’ as an even number.

Model’s Generated Rationale (Incorrect)
"Trong trường hợp này, biển số P.131c áp dụng vì nó cấm đỗ xe vào những ngày chẵn. Ngày 17/5/2025 là ngày chẵn, do đó biển báo này có hiệu lực." **English Translation:** *"In this case, sign P.131c applies because it prohibits parking on even days. May 17, 2025 is an even day, therefore this sign is in effect."*

6 Conclusion

In this paper, we presented our system for the VLSP 2025 MLQA-TSR shared task, detailing a comprehensive, two-stage approach for multimodal legal information retrieval and visual question answering. Our retrieval system’s core innovation is a rule-based heterogeneous knowledge graph that accurately models the relationships between legal articles, traffic signs, and tabular data. This structured approach, combined with a multimodal reranker and an adaptive dynamic top-k selection strategy, proved highly effective, securing Rank 2 in the Legal Document Retrieval task with an F2-score of 0.611. For the VQA task, our three-stage reasoning pipeline achieved an accuracy of 0.623, placing us 8th overall.

Acknowledgements

This research is funded by Vietnam National University HoChiMinh City (VNU-HCM) under grant number DS.C2025-26-10.

References

Jina AI. 2024. jina-reranker-m0: Multilingual multimodal document reranker. https://huggingface.co/jinaai/jina-reranker-m0. Accessed: 2025-10-03.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

Muhammad Rafsan Kabir, Rafeed Mohammad Sultan, Fuad Rahman, Mohammad Ruhul Amin, Sifat Momen, Nabeel Mohammed, and Shafin Rahman. 2025. Legalrag: A hybrid rag system for multilingual legal information retrieval. *Preprint*, arXiv:2504.16121.

Zijian Li, Qingyan Guo, Jiawei Shao, Lei Song, Jiang Bian, Jun Zhang, and Rui Wang. 2024. Graph neural network enhanced retrieval for question answering of llms. *Preprint*, arXiv:2406.06572.

Yanran Tang, Ruihong Qiu, Yilun Liu, Xue Li, and Zi Huang. 2023. Casegnn: Graph neural networks for legal case retrieval with text-attributed graphs. *Preprint*, arXiv:2312.11229.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xu Yuan, Liangbo Ning, Wenqi Fan, and Qing Li. 2025. mkg-rag: Multimodal knowledge graph-enhanced rag for visual question answering. *Preprint*, arXiv:2508.05318.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Zhengyuan Zhu, Daniel Lee, Hong Zhang, Sai Sree Harsha, Loic Feujio, Akash Maharaj, and Yunyao Li. 2025b. Murar: A simple and effective multimodal retrieval and answer refinement framework for multimodal question answering. *Preprint*, arXiv:2408.08521.