VLSP 2025 MLQA-TSR Challenge: Vietnamese Multimodal Legal Question Answering on Traffic Sign Regulation

Son T. Luu[1,2,3], Trung Vo[1], Hiep Nguyen[1], Khanh Quoc Tran[2,3], Kiet Van Nguyen[2,3], Vu Tran[1], Ngan Luu-Thuy Nguyen[2,3], Le-Minh Nguyen[1],
[1]Japan Advanced Institute of Science and Technology, Ishikawa, Japan,
[2]University of Information Technology, Ho Chi Minh City, Vietnam,
[3]Vietnam National University, Ho Chi Minh City, Vietnam
Correspondence: sonlt@jaist.ac.jp, ngannlt@uit.edu.vn, nguyenml@jaist.ac.jp

Abstract

This paper presents the VLSP 2025 MLQA-TSR - the multimodal legal question answering on traffic sign regulation shared task at VLSP 2025. VLSP 2025 MLQA-TSR comprises two subtasks: multimodal legal retrieval and multimodal question answering. The goal is to advance research on Vietnamese multimodal legal text processing and to provide a benchmark dataset for building and evaluating intelligent systems in multimodal legal domains, with a focus on traffic sign regulation in Vietnam. The best-reported results on VLSP 2025 MLQA-TSR are an F2 score of 64.55% for multimodal legal retrieval and an accuracy of 86.30% for multimodal question answering.

1 Introduction

Multimodal Question Answering (QA) is a challenging task in Natural Language Processing (NLP) that requires systems to understand and integrate multiple data types—such as text and images—to extract the correct information. Multimodal QA plays an important role in building intelligent systems because it offers a natural, user-friendly way for humans to search for information (Luu-Thuy Nguyen et al., 2023). Moreover, legal text processing is also difficult for NLP: legal language is highly formal, structurally complex, and rich in specialized terminology that presupposes substantial knowledge of legal concepts and principles (Anh et al., 2023). To answer legal questions correctly, systems must not only have an in-depth understanding of legal documents but also perform reasoning over legal information to arrive at the correct conclusions.

In recent years, many competitions have been organized on legal text processing to boost research in artificial intelligence for legal text processing. The COLIEE (Goebel et al., 2024) is an annual competition about legal text processing that targets the in-depth legal text understanding tasks like legal entailment and legal question answering in the English language. Similar to the COLIEE series, the ALQAC (Do et al., 2024) and VLSP-2023-LTER (Tran et al., 2024) are two shared tasks about legal text processing, including legal retrieval, legal question-answering, and legal textual entailment in Vietnamese legal documents. These competitions provide valuable benchmark datasets and a forum for research attempts in legal processing. However, these competitions focus on text only for legal domains, which motivates us to construct a multimodal competition about legal processing.

Building on previous Vietnamese legal NLP competitions such as ALQAC (Do et al., 2024), **VLSP 2025 MLQA-TSR** (VLSP 2025 **M**ultimodal **L**egal **Q**uestion **A**nswering on **T**raffic **S**ign **R**egulation) introduces a multimodal task centered on traffic sign regulation. As illustrated in Figure 1, answering a legal question requires understanding both the textual question and the accompanying image of traffic signs and then consulting the relevant statutes on road traffic and safety (VIETNAM, 2024) as well as the *Regulation on Traffic Signs and Signals* (Ministry of Transport, 2024). VLSP 2025 MLQA-TSR aims to spur the development of intelligent systems that can interpret multimodal legal inputs and retrieve correct answers to users' questions. The challenge comprises two subtasks: *Legal Retrieval* (SubTask 1) and *Legal Question Answering* (SubTask 2), and is hosted on the CodaBench platform (Xu et al., 2022).

Overall, this paper provides an overview of the VLSP 2025 MLQA-TSR share task. We summarize two main contributions in this paper as:

- First, we provide a benchmark dataset about multimodal legal question answering on traffic sign regulation in Vietnam. The dataset consists of two multimodal tasks: legal retrieval and legal question answering.
- Second, we organize the shared task VLSP

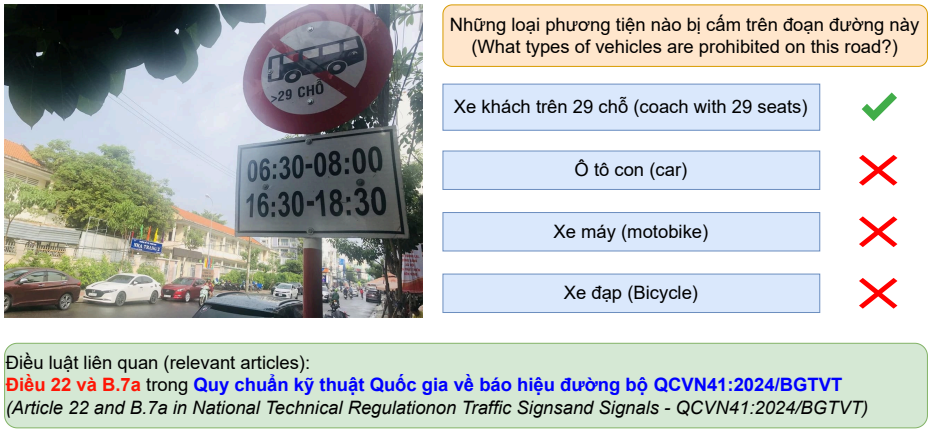


Figure 1: A sample of a legal question about a traffic sign.

2025 MLQA-TSR at VLSP 2025. We obtained 13 team submissions for SubTask 1 and 19 team submissions for SubTask 2.

The remainder of the paper is organized as follows. Section 2 describes the two subtasks in detail. Section 3 presents the data construction process and an overall analysis. Section 4 summarizes participant approaches and the baseline used in the competition. Section 5 reports the results and rankings. Finally, Section 6 concludes the paper and outlines future directions.

2 Task description

VLSP 2025 MLQA-TSR shared task focuses on enhancing the ability of computers to understand legal text in multimodal scenarios about traffic sign regulation. The shared task includes two subtasks: multimodal legal retrieval (Subtask 1) and multimodal legal question answering (Subtask 2).

2.1 Subtask 1: Multimodal Legal Retrieval

Task definition: Given a multimodal question $q = (q_{\text{text}}, q_{\text{image}})$ consisting of two components, where q_{text} is the question in textual form and q_{image} is the query image corresponding to q_{text}—and a law database $\mathcal{D} = \{d_i \mid i = 1, \ldots, n\}$ containing articles from legal documents, the goal is to determine a ranked list of relevant articles $\mathcal{R} = \{d_i \mid d_i \in \mathcal{D}, i = 1, \ldots, k, k \leq n\}$ for the question q. Each article d_i may be text-only or multimodal (e.g., text, images, and/or tables).

Evaluation metric: We use the F2-score for assessing the performance of the retrieval model since the F2-score places more in recall, which is concerned about false negatives more than false positives. The F2-score for a question q is computed as Equation 3. The final F2-score is determined by averaging the F2-score over the evaluation sets.

$$Precision_q = \frac{\#correct\_retrieved\_articles}{\#total\_retrieved\_articles} \quad (1)$$

$$Recall_q = \frac{\#correct\_retrieved\_articles}{\#total\_relevant\_articles} \quad (2)$$

$$F2_q = 5 * \frac{Precision_q * Recall_q}{4 * Precision_q + Recall_q} \quad (3)$$

2.2 Subtask 2: Multimodal Legal Question Answering

Task definition: Given a multimodal question $q = (q_{\text{text}}, q_{\text{image}})$—with q_{text} as the textual query and q_{image} as the corresponding image—and a list of relevant articles $\mathcal{R} = \{d_i \mid d_i \in \mathcal{D}, i = 1, \ldots, k, k \leq n\}$ for q, the objective is to predict the correct answer a from four multiple-choice options for the multimodal question q.

Evaluation metric: Subtask 2 is formulated as multiple-choice question answering. Therefore, we employ *Accuracy* as the primary metric, measuring the proportion of correctly predicted answers over the evaluation set. The accuracy is determined as Equation 4

$$Accuracy = \frac{\#total\_correct\_answers}{\#total\_questions} \quad (4)$$

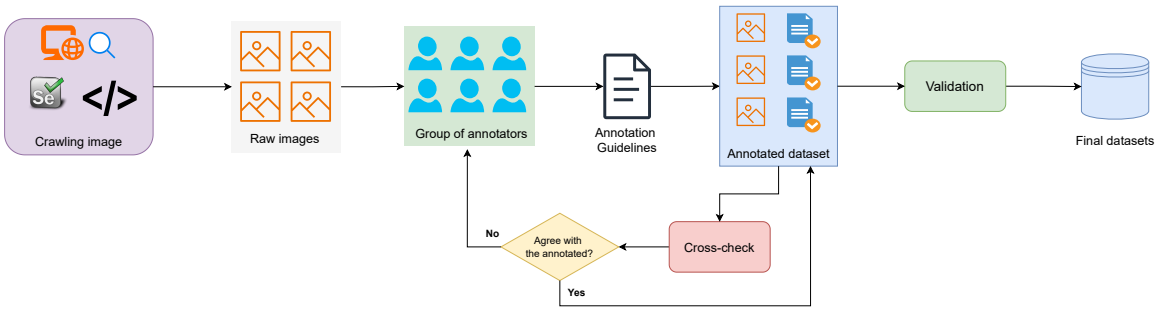


Figure 2: Data Creation Process.

3 Dataset

3.1 Data Construction

Figure 2 shows the overview of the data construction process for the VLSP 2025 MLQA-TSR share task. The process consists of three main stages as described follows:

- **Stage 1 - Data Collection:** We search the images of traffic signs on streets in Vietnam on Google search, then we collect them automatically by using the Selenium tool based on HTML processing. After stage 1, we have a raw image set that contains various traffic signs on the street, and also non-relevant images to traffic signs. We then manually remove the images that are not relevant to traffic signs.

- **Stage 2 - Data Annotation:** We hire a group of 8 annotators who are undergraduate students and give them the annotation guidelines. After reading the annotation guidelines, we let the annotator create a question and an answer based on the image of traffic signs. First, annotators are required to read the Regulation on Traffic Signs and Signals and the Road Traffic and Safety Law, then give a list of the most relevant articles to the question. Second, the annotators make the correct answer according to the question. There are two kinds of questions: multiple-choice and yes/no questions. For the multiple-choice, annotators are required to create four different choices and mark the correct one. The choices of Yes/No questions have only two options, including "Đúng" indicating "Yes" and "Sai" representing "No".

- **Stage 3 - Cross-checking:** We let the annotators perform cross-checking among the annotated data. One will check the correctness of the answers and relevant articles to the question and traffic sign images, syntax, and typos of the questions and answers of others. If the annotator disagrees with an annotated sample, the disagreement sample will be sent back to the group of annotators for re-annotating.

- **Stage 4 - Validation:** We let the annotators perform final validation of annotated data. The main criteria for validation include: the consistency between the question and the traffic signs, the correctness of the relevant articles and the answers, and the typos and syntax in the question and answers. If any samples do not satisfy the criteria, we remove them from the dataset.

The final dataset is split into three sets: training, public test, and private test sets to serve for the shared task. Besides, we also provide the law database, including the articles in both the National Regulation on Traffic Signs and Signals (**QCVN 41:2024/BGTVT**) and the Law on Road Traffic Order and Safety (**36/2024/QH15**). In the National Regulation on Traffic Signs and Signals, we represent the image in the articles with a format *«IMAGE: image_file.jpg /IMAGE»* and the table as the following format *«TABLE: table_html_code /TABLE»*. The law database is provided to the participants as a JSON file format along with a directory containing corresponding images.

3.2 Data Analysis

Table 1 summarizes the overall information about the two legal documents used in the law database. It can be seen that the document **QCVN 41:2024/BGTVT** - "National Regulation on Traffic Signs and Signals" contains both image and table data in the article, while the **36/2024/QH15** - "Law on Road Traffic Order and Safety" only has text in

the document. Also, the number of articles in the National Regulation on Traffic Signs and Signals is significantly more than the Law on Road Traffic Order and Safety, since the National Regulation on Traffic Signs and Signals contains a detailed description of the technical specifications and the meaning of various traffic signs in road traffic in Vietnam.

Next, Table 2 illustrates the overall statistics about the three sets used in the VLSP 2025 - MLQA-TSR. In the training and public test sets, the proportion between multiple-choice and Yes/No questions is 6/3, while this proportion is 5-5 in the private test to ensure the objective performance of the question-answering models for the type of question. Also, the average length of a question in the private test set is higher than in the training and public test sets, challenging the generability of question answering models in generating correct answers for the questions (The length of the question is computed according to the number of tokens in the question. We segment the question text into token-level by using the Pyvi[1]).

Table 1: Overview information about two legal documents in the law database

Law ID	**QCVN 41:2024/BGTVT**	**36/2024/QH15**
Law Name	National Regulation on Traffic Signs and Signals.	Law on Road Traffic Order and Safety.
# Articles	313	89
# Image	761	0
# Table	212	0

Table 2: Statistics about the provided dataset in VLSP 2025 MLQA-TSR

	Train	**Public Test**	**Private Test**
# Total questions	530	100	146
# Total images	304	90	104
# Multiple choices questions	376	65	74
# Yes/No questions	154	35	72
Max length of question	69	42	76
Min length of question	5	5	8
Average length of question	16.95	14.96	27.00
Max relevant articles	8	7	10
Min relevant articles	1	1	1
Average relevant articles	2.31	2.19	2.76

On the other hand, the number of relevant articles per question on the three sets is similar, with around two articles for a question. As shown in Figure 3, most questions in the datasets have two relevant articles. For the training and public test sets, the number of relevant articles usually falls into 1 to 2 articles, while it often falls into about 2 to 4 articles in the private test. Overall, the distribution of data in the private test is slightly different from the training and public test sets to ensure the objective of the model and avoid overfitting.

[1] https://pypi.org/project/pyvi/

4 Method

4.1 Baseline Methods

We adopt **BGE Visualized** (Zhou et al., 2024) as a strong and efficient baseline for multimodal retrieval. We encode each article in the law database into a d-dimensional embedding (we use d=1024), yielding a matrix of shape (K, d), where K is the number of articles. Given a query consisting of the textual question and the associated image, we encode it into a single d-dimensional embedding and compute dot-product similarities against all article embeddings. We then retrieve the *top-5* most similar articles as candidates for the query.

For QA, we use **Vintern** (Doan et al., 2024), a Vietnamese Multimodal Large Language Model (MLLM). We prompt the `Vintern-3B-beta` checkpoint to generate answers. For the two question types (multiple choice and Yes/No), we employ the following prompt templates:

Multiple-choice questions

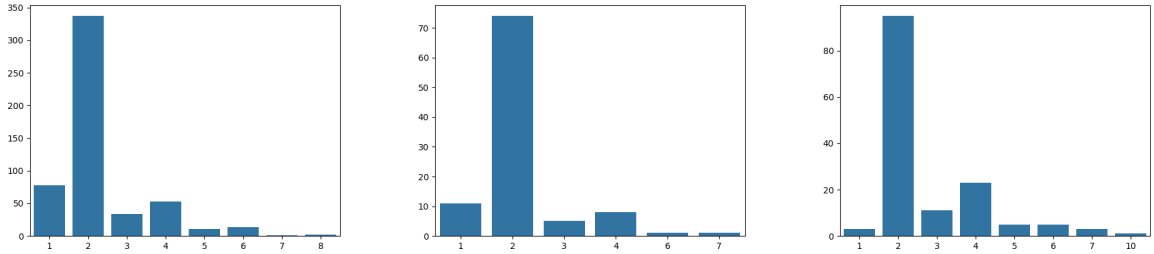
<image>
<question>
A. Lựa chọn 1. (*English*: Choice 1)
B. Lựa chọn 2. (*English*: Choice 2)
C. Lựa chọn 3. (*English*: Choice 3)
D. Lựa chọn 4. (*English*: Choice 4)

Trả lời bằng một trong bốn đáp án: A, B, C hoặc D. Không giải thích thêm
(*English*: Answer by one of four choices: A, B, C, or D without explanation)

Yes/No questions

<image>
<question>

Trả lời bằng một trong hai đáp án: Đúng hoặc Sai. Không giải thích thêm.
(*English*: Answer by one of two values: Yes or No without explanation)



(a) Training set (b) Public Test (c) Private Test

Figure 3: Distribution of relevant articles in three sets

4.2 Proposed Methods by Participants

This section summarizes the methodologies proposed by the top–5 teams for each subtask in VLSP 2025 MLQA-TSR. Participants were required to use only open-source LLMs whose code and model weights are publicly available (e.g., GitHub, Hugging Face); commercial LLMs (e.g., ChatGPT, Gemini, Claude) were not allowed. Additionally, the use of any external data beyond the competition resources was prohibited. Below are the reported approaches:

- **SmartbotIC:** The team uses Qwen2.5-VL and InternVL3 with zero-shot prompting to generate answers from the input question and image, together with the retrieved articles. To improve efficiency and quality, they apply several preprocessing steps, including concatenating multiple images into a single image and converting HTML tables to Markdown to reduce input length.

- **Berry:** For retrieval, the team encodes questions and choices with Jina Embeddings (text) and uses C-RADIOv2-B to encode images, while OWLv2 detects traffic-sign objects whose features are also embedded. All vectors are stored in a Qdrant vector database; subtask 1 results are obtained via vector search to return the top-k most similar candidates. For QA, they apply few-shot prompting with examples retrieved from the vector store and generate final answers using the Llama 4 Maverick.

- **Tanka_CDS:** The team preprocesses the law database by converting HTML tables to normalized text, chunking, and embedding articles with CLIP. For retrieval, YOLOv8n is fine-tuned to extract key regions in traffic-sign images as patches; LLaMA3.2-Vision then encodes the question with the corresponding patches. For QA, LLaMA3.2-Vision generates answers based on the question, image, and the retrieved articles.

- **chmod+x:** The team models a *heterogeneous graph* to represent relationships among images, text, and tables in the law database. They then apply the Jina Reranker to re-rank visual documents and perform graph matching combined with similarity search to retrieve the top-k candidates for a given question and image. They further propose a dynamic top-k filtering mechanism based on counting relevant traffic signs to adapt the candidate set size to question complexity. For QA, they use an ensemble of two multimodal LLMs: Qwen2.5-VL-7B and InternVL3-8B.

- **Metamorphic:** The team designs a systematic workflow to produce final answers from user inputs (question, image, choices). They construct graph representations for images (*ImageSubGraph*) and for articles (*ArticleSubGraph*) to retrieve enriched information (e.g., scene descriptions and fine-grained traffic-law content). This information is merged into a unified prompt to guide the LLM. In the framework, Gemma2-27B and DeepSeek are used to encode and process multimodal inputs for QA. YOLO is also employed for traffic-sign detection as an additional feature.

- **LifeIsTough:** The team preprocesses the law database by converting HTML tables to Markdown and normalizing text. For database images, Gemma-3-12B is used to crop regions so that only traffic signs are retained. Text and images are then encoded with CLIP

and stored in Qdrant. For retrieval, the team first applies YOLOE to detect traffic signs in the input image, then validates the detections against the input question using Gemma-3-12B. The enriched query is encoded with CLIP to construct a query vector, which is used to search the law database (already embedded) for relevant articles. For QA, the retrieved knowledge is combined with the question in the prompt, and Gemma-3-12B generates the final answer.

- **TechNova:** For retrieval, the team implements a two-branch architecture. The first branch generates separate image and text embeddings for the entire training set. The second branch builds a dense corpus retriever by indexing text-chunk embeddings from the full law database in a FAISS index. To retrieve relevant articles, the query (image + question) is fused into a multimodal embedding and compared against article embeddings. The team uses gme-Qwen2-VL-2B-Instruct to produce retrieval embeddings. For QA, they employ Qwen2.5-VL-72B-Instruct with Chain-of-Thought prompting: first describe the visual scene, then apply the legal context, and finally reason to a conclusion.

Overall, the Multimodal Legal Retrieval Task (Subtask 1) relies on constructing a contextual multimodal embedding space to perform similarity search or re-ranking. Robust multimodal embedding models such as CLIP (Radford et al., 2021), Jina Embedding (Günther et al., 2025), C-RADIOv2 (Heinrich et al., 2025) and multimodal LLMs like Gemma-3 (Team et al., 2025) and Qwen2-VL (Wang et al., 2024). To enhance the accuracy of the retrieval task, participants often employ traffic sign detection models like OWL (Minderer et al., 2023) or YOLO (Yaseen, 2024) to filter the key information about traffic signs in the images or crop them by Gemma-3 (Team et al., 2025). In addition, data preprocessing steps such as text normalization, image filtering, combination, concatenation, and HTML table transformation are also frequently used to enhance the performance of the model in vector embedding. In addition, the vector databases, such as Qdrant [2] or FAISS [3], are usually used in a retrieval task to serve for vector-space similarity searching. Moreover, several teams use retrieval and searching on graphs to improve the performance of the retrieval task by efficiently representing the multimodal data in graphs to capture semantic relation information. For the Multimodal Legal Question Answering Task (Subtask 2), the vision LLMs like Qwen2.5-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), Gemma2 (Team et al., 2024), and LLaMa3.2-Vision (Grattafiori et al., 2024) are used by almost all participants, indicating the robustness and efficiency of vision LLMs for multimodal QA. Zero-shot prompting, Chain-of-Thought (Wei et al., 2022), and few-shot prompting are mostly used techniques by participants to instruct vision LLMs in generating the answer.

5 Results

Table 3 shows the ranking results of participants for Subtask 1 on the private test. The top 1 team - **LifeIsTough**, with the efficient retrieval method that filters the key features in the traffic sign image via cropping by LLM and latent vector embedding construction by CLIP (Radford et al., 2021), achieves the highest results with 64.55% by F2 score. The **chmod+x** team is runner-up with 61.13%, and **TechNoVa** places third with 59.91% by F2 score. Additionally, there is a significant gap between the top 5 teams with others in Subtask 1, where the top 5 teams attain a performance by F2 score of more than 50%, and others obtain lower than 50% of performance by F2 score. All teams in Subtask 1 obtain performance better than the baseline methods, indicating the efficiency of the proposed method for the legal retrieval task. Since the highest results for the legal retrieval task are approximately 65%, there is still room for further improvement in this task.

Next, Table 4 illustrates the ranking of participants for Subtask 2 on the private test[4]. The top 1 team - **Berry** obtains optimistic results for this task with 86.30% by Accuracy by employing the LLama4-Maverick with efficient few-shot prompting. **SmartbotIC** is the runner-up team with 83.56%, and **TechNova** obtains 3^{rd} rank with 78.08% by Accuracy. Overall, it can be seen that the gap between the top 5 teams with others in Subtask 2 is not as much as in Subtask 1, indicating the efficiency and robustness of the proposed methodologies by participants for the multimodal legal question answering.

[2] https://qdrant.tech/

[3] https://github.com/facebookresearch/faiss

[4] We report the results that performance over the baseline

Table 3: Results for subtask 1 - Multimodal Legal Retrieval

Team name	F2 score	Rank
LifeIsTough	**0.6455361395**	**1**
chmod+x	**0.6113748745**	**2**
TechNova	**0.5991697165**	**3**
SmartbotIC	**0.5790135683**	**4**
Berry	**0.5432150682**	**5**
DHDD	0.4511688070	6
Tanka_CDS	0.2459355607	7
AIO_VNM	0.2384666964	8
MealsRetrieval	0.1548250424	9
OpenCubee	0.1533444945	10
Come4Win	0.1413907401	11
LexTraffic	0.1358808937	12
BASELINE	0.1276493485	13

Table 4: Results for subtask 2 - Multimodal Legal Question Answering

Team name	Accuracy score	Rank
Berry	**0.8630136986**	**1**
SmartbotIC	**0.8356164384**	**2**
TechNova	**0.7808219178**	**3**
Tanka_CDS	**0.7328767123**	**4**
Metamorphic	**0.7260273973**	**5**
Hallucinators	0.7123287671	6
LifeIsTough	0.6712328767	7
chmod+x	0.6232876712	8
OpenCubee	0.6095890411	9
LexTraffic	0.5958904110	10
AIO_VNM	0.5684931507	11
NaN	0.5616438356	12
SoftMind_AIO	0.5000000000	13
BASELINE	0.4520547945	14

In comparison with ALQAC 2024 (Do et al., 2024) in Vietnamese language and COLIEE 2024 (Goebel et al., 2024) in English language - the two competitions about legal document processing, it can be seen that the best performance on the legal retrieval task is about 87% by F2 score at ALQAC 2024, and 44% by F2 score at COLIEE 2024 (Task 1), while the performance of legal question answering task is significantly higher with 98% by Accuracy at ALQAC 2024 and 82% by Accuracy at COLIEE 2024 (Task 4). In general, the legal retrieval task is more challenging than question answering, since legal documents have a complex structure and specialized legal terminologies that require an in-depth understanding between the users' queries and the legal documents to extract correct and valuable information. In the scenario of multimodal, the retrieval system not only focuses on text but is also concerned about the latent information from the image to provide the correct answer.

6 Conclusion

This paper introduces VLSP 2025—MLQA-TSR, a new multimodal shared task in legal text processing designed to advance research on low-resource languages, with a primary focus on Vietnamese. The task comprises two subtasks: (1) multimodal legal retrieval and (2) multimodal question answering. The best-performing systems achieve an F2 score of 64.55% on the multimodal legal retrieval subtask and an accuracy of 86.30% on the multimodal question answering subtask. VLSP 2025—MLQA-TSR attracted a range of innovative methodologies leveraging state-of-the-art models across both subtasks, offering valuable momentum for research in Vietnamese multimodal legal text processing. Finally, VLSP 2025—MLQA-TSR provides a benchmark dataset for building and evaluating intelligent systems in the legal domain and multimodal tasks in Vietnamese, specifically centered on traffic sign regulation. The dataset and baseline code are published at https://github.com/sonlam1102/VLSP2025-MLQA-TSR.

References

Dang Hoang Anh, Dinh-Truong Do, Vu Tran, and Nguyen Le Minh. 2023. The impact of large language modeling on natural language processing in legal texts: A comprehensive survey. In *2023 15th International Conference on Knowledge and Systems Engineering (KSE)*, pages 1–7.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Dinh-Truong Do, Son T. Luu, Trang Pham, Trung Vo, Nguyen-Hoang Chu, Quang-Huy Chu, Cuong Nguyen, Minh Nguyen, An Trieu, Dat Nguyen, Thanh Tran, Cong Nguyen, Hiep Nguyen, Chau Nguyen, Nguyen-Khang Le, Dieu-Hien Nguyen, Binh Dang, Phuong Nguyen, Ha-Thanh Nguyen, and 2 others. 2024. A summary of the alqac 2024 competition. In *2024 16th International Conference on Knowledge and System Engineering (KSE)*, pages 422–427.

Khang T. Doan, Bao G. Huynh, Dung T. Hoang, Thuc D. Pham, Nhat H. Pham, Quan T. M. Nguyen, Bang Q.

Vo, and Suong N. Hoang. 2024. Vintern-1b: An efficient multimodal large language model for vietnamese.

Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2024. Overview of benchmark datasets and methods for the legal information extraction/entailment competition (coliee) 2024. In *JSAI International Symposium on Artificial Intelligence*, pages 109–124. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and Han Xiao. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval.

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. 2025. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22487–22497.

Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong T.D. Vo, Khanh Quoc Tran, and Kiet Van Nguyen. 2023. Evjvqa challenge: Multilingual visual question answering. *Journal of Computer Science and Cybernetics*, 39(3):237–259.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36:72983–73007.

MoT Ministry of Transport. 2024. National technical regulation on traffic signs and signals.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Vu Tran, Ha-Thanh Nguyen, Trung Vo, Son T Luu, Hoang-Anh Dang, Ngoc-Cam Le, Thi-Thuy Le, Minh-Tien Nguyen, Truong-Son Nguyen, and Le-Minh Nguyen. 2024. Vlsp 2023–lter: A summary of the challenge on legal textual entailment recognition. *arXiv preprint arXiv:2403.03435*.

NATIONAL ASSEMBLY OF VIETNAM. 2024. Law on road traffic order and safety.

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

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

Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. *Patterns*, 3(7):100543.

Muhammad Yaseen. 2024. What is yolov8: An in-depth exploration of the internal features of the next-generation object detector. *arXiv preprint arXiv:2409.07813*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.