LexiSignVQA: A Unified Training-free Multi-stage Approach to Multimodal Legal Question Answering on Traffic Sign Rules

Phung Xuan Pham [*]
Vietnam Silicon
phung.pham@vnsilicon.net

Duc Quang Le [*]
Viettel AI, Viettel Group
duclq3@viettel.com.vn

Tuan Hau Tran [*]
Viettel AI, Viettel Group
tuanth22@viettel.com.vn

Thinh Nguyen-Truong Huynh [*] [†]
Viettel AI, Viettel Group
thinhhnt@viettel.com.vn

Abstract

The task of multimodal legal question answering on traffic sign rules (MLQA-TSR) presents unique challenges due to the need for jointly interpreting visual and textual information in regulatory contexts. In this paper, we propose **LexiSignVQA**, a unified, training-free, multi-stage approach developed for the VLSP 2025 MLQA-TSR shared task. Our approach integrates traffic sign detection, image embedding, and vision–language modeling with a structured preprocessing procedure that aligns traffic sign images with their corresponding legal provisions. By combining simple yet effective image processing for clean legal databases with traffic sign detection models for real-world scenarios, our method achieves both efficiency and robustness. Experimental results on the MLQA-TSR dataset demonstrate that **LexiSignVQA** ranked first in multimodal retrieval (Subtask 1) and seventh in legal question answering (Subtask 2). Furthermore, our analysis reveals the complementary strengths of conventional segmentation versus learning-based detection and highlights the role of embeddings in addressing directional reasoning in traffic signs. These findings underscore the potential of hybrid, training-free frameworks for advancing multimodal legal reasoning and practical applications in traffic law compliance. The source code is available at https://github.com/phungpx/LexiSignVQA.

1 Introduction

The task of visual question answering (VQA) (Abacha et al., 2019) has long been recognized as a central challenge in artificial intelligence, particularly within the field of natural language processing (NLP). In recent years, research on domain-specific VQA systems has gained considerable attention (Farea and Emmert-Streib, 2025; Agrawal et al., 2015), as such systems can provide practical solutions in specialized contexts. Among these, legal VQA (Le et al., 2024) is especially significant due to the growing demand for tools that assist users in retrieving, interpreting, and applying legal information in real-world scenarios.

Within the domain of road traffic safety, strict adherence to regulations is essential for protecting both human life and property. Traffic signs, as fundamental carriers of legal and safety information, play a vital role in ensuring compliance with the law. Correct interpretation of these signs not only facilitates safer road usage but also enhances public awareness of traffic rules and their legal implications. However, the multimodal nature of traffic sign rules—which combine visual and textual components—poses unique challenges for the design of robust VQA systems (Jabri et al., 2022).

To address these challenges, the VLSP 2025 Multimodal Legal Question Answering on Traffic Sign Rules (MLQA-TSR) shared task has been introduced (The Association for Vietnamese Language and Speech Processing VLSP, 2025). This task aims to advance research at the intersection of NLP and multimodal learning by integrating both textual legal documents and visual traffic sign data. The ultimate goal is to develop intelligent systems capable of supporting users in understanding traffic sign meanings and their corresponding legal provisions.

The competition is structured into two subtasks. The first subtask focuses on multimodal retrieval, where participants must retrieve relevant legal articles from the Vietnamese Law on Road Traffic Order and Safety and the National Technical Regulation on Traffic Signs and Signals, given a natural language query accompanied by a road scene image. The second subtask extends this problem to question answering, where participants must not only identify relevant legal articles but also provide direct answers in either multiple-choice or Yes/No

[*] These authors contributed equally to this work.
[†] Corresponding author

format. This dual-task structure highlights the need for models that effectively integrate multimodal data while remaining robust in handling practical legal VQA scenarios.

One major challenge of this task lies in their limited ability to accurately interpret directional information, such as left, right or straight-ahead indicators on traffic signs. This weakness often leads to misalignment between visual cues and textual reasoning, resulting in errors in downstream tasks such as answering legal questions or navigation. Addressing this issue requires capturing the spatial and semantic nuances of directional symbols.

In this work, we present our approach, **LexiSign-VQA**, developed for the MLQA-TSR challenge. Our method integrates traffic sign detection, vision–language modeling, and image embedding into a unified multi-stage pipeline designed to address both subtasks. By combining conventional rule-based image processing with state-of-the-art deep learning methods, our framework demonstrates competitive performance while maintaining efficiency and reproducibility.

The main contributions of this work can be summarized as follows:

- We propose a unified, training-free, multi-stage approach that integrates traffic sign detection, image embedding, and vision–language modeling to address the dual challenges of multimodal retrieval and question answering in the legal domain.

- We designed an efficient preprocessing procedure that transforms raw legal articles into a structured format and segments road scene images into unique traffic sign patches, thereby establishing a direct alignment between textual legal provisions and their corresponding visual representations, which helps the model understand directions and provides insights for vision–language models.

- Through extensive experiments on the VLSP 2025 MLQA-TSR dataset, we demonstrate that our approach achieves competitive results, ranking first and seventh on the subtasks 1 and 2, respectively. Furthermore, our analysis highlights the relative strengths of conventional versus learning-based detection methods and the effectiveness of embedding models for traffic sign retrieval.

	Subtask 1	Subtask 2
No. train samples	530	530
No. public test samples	50	50
No. private test samples	146	46

Table 1: Overview statistics of the dataset

2 Task Definition

In the context of road traffic safety, ensuring strict adherence to regulations is vital for the protection of human life and property. A critical component of this process is the correct interpretation and compliance with traffic signs and signals, which serve as the foundation for safe road usage. This task uniquely combines textual and visual modalities, aiming to create systems capable of assisting users in comprehending traffic sign meanings and their legal implications, thereby raising public awareness of traffic safety. The shared task consists of two subtasks:

2.1 Subtask 1: Multimodal Retrieval

Participants are required to retrieve relevant legal articles from the Vietnamese traffic law (LawDB) consisting of the *Law on Road Traffic Order and Safety (36/2024/QH15; 313 articles)* and the *National Technical Regulation on Traffic Signs and Signals (QCVN 41:2024/BGTVT; 89 articles)*, given a natural language query and an accompanying image of the real-world road setting scenario. Figure 1 illustrates the distribution of relevant articles per question, indicating that questions with two relevant articles constitute the majority of Subtask 1's dataset.

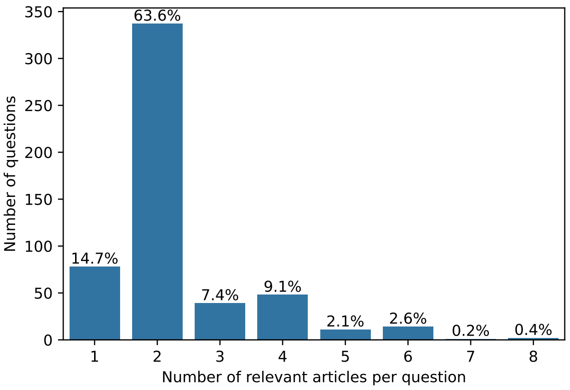


Figure 1: Distribution of relevant articles per question for Subtask 1

2.2 Subtask 2: Question Answering

Building on the outputs of Subtask 1, participants must provide answers in either multiple-choice or Yes/No formats. Each question is posed in natural language, supplemented by a traffic sign image and reference to specific legal provisions. This dual structure not only emphasizes the integration of multimodal data but also encourages the development of robust methodologies capable of addressing practical challenges in legal information retrieval and question answering. Figure 2 demonstrates that the training dataset is dominated by multiple-choice questions, whereas the test sets exhibit a more balanced mix of question types.

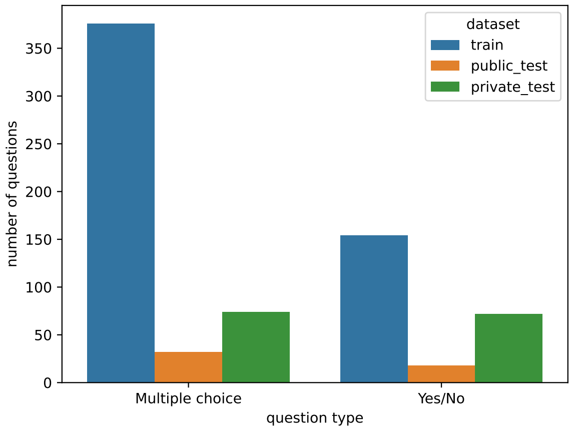


Figure 2: Distribution of question types for Subtask 2

3 Methodology

This section outlines the integration of existing models (e.g., traffic sign detection, vision-language, and image embedding systems) into a sophisticated yet efficient multi-stage pipeline. The framework is designed to address both tasks by employing lightweight preprocessing, generating efficient image embeddings, and applying traffic sign detection alongside retrieval algorithms. Figure 4 presents the overall pipeline of the proposed approach for Subtasks 1 and 2.

3.1 Preprocessing

To ensure consistent and efficient downstream processing, both textual and visual data undergo a series of preprocessing steps, illustrated in Figure 3. Algorithm 1 presents the overall preprocessing workflow.

Algorithm 1 Preprocessing for each raw article input

```
Require: Legal Text $T$, Images $\{I_1, I_2, \dots, I_n\}$ of Law ID $L$ and Article ID $A$
 Preprocess $T$ to remove styles and convert to Markdown
 Crop $I_i$ into sub-images $\{I_{i1}, I_{i2}, \dots, I_{im}\}$ using heuristic segmentation
 for each $I_k$ in $\bigcup_{i=1}^{n} \{I_{i1}, \dots, I_{im}\}$ do
     $signname_k \leftarrow$ ExtractSignInfo$(I_k, T)$
     Generate embedding $E_k \leftarrow$ Embed$(I_k)$
     Store $(E_k, L, A, signname_k)$ in Qdrant
 end for
```

Legal text preprocessing: Raw legal articles from LawDB, formatted as HTML tables, are first converted to Markdown. During this conversion, stylistic elements (e.g., `class="MsoNormalTable"`, `style="width:100.0%;..."`) are ignored in order to reduce redundant text and minimize the number of tokens required for subsequent processing. This step ensures that the content remains semantically faithful to the original legal description while being computationally efficient.

Traffic sign patch splitting: Since a single raw image can contain multiple traffic signs, we first apply a heuristic image processing algorithm to segment it into distinct regions, each corresponding to a unique sign. Algorithm 2 presents a traditional image processing technique for extracting traffic signs from clean images with uniform white backgrounds.

Traffic sign information extraction Subsequently, a vision–language model (VLM) (Team et al., 2025) is employed to align these images with the preprocessed legal text. This allows the model to generate both a concise title and a semantically faithful description for each traffic sign, thereby establishing a direct link between visual input and regulatory context.

Embedding generation and storage: Following preprocessing, the images are transformed into latent representations using an image embedding model (**?**; Tschannen et al., 2025). These embeddings are indexed in a vector database (Han et al., 2023), where each entry is enriched with a payload consisting of the corresponding law ID, article ID, and sign name. Such a structured storage scheme not only ensures efficient retrieval but also establishes the foundation for downstream applications, most notably traffic sign rule question answering.

Algorithm 2 Conventional Image Processing for Traffic Sign Extraction in LawDB

```
Require: Traffic sign image $I_i$
Ensure: Cropped traffic sign patches $\{I_{i1}, I_{i2}, \ldots, I_{im}\}$
Add border of size $b$ to $I$ to prevent edge loss
Convert $I$ to grayscale $\rightarrow I_{i,gray}$
Apply Gaussian blur on $I_{i,gray} \rightarrow I_{i,blur}$
Apply binary inverse thresholding on
    $I_{i,blur} \rightarrow I_{i,thresh}$
Extract external contours
    $\mathcal{C} \leftarrow \text{Contours}(I_{i,thresh})$
Filter $\mathcal{C}$ by:
    • $\text{Area}(c) \geq \alpha \cdot \text{Area}(I)$    ▷ minimum contour area ratio
    • $\text{Width}(c) > 10$ and $\text{Height}(c) > 10$
    • Ignore $c$ when $\frac{\text{Width}(c)}{\text{Height}(c)} \leq 4$
Compute bounding boxes
    $\{B_{i1}, B_{i2}, \ldots, B_{im}\}$ for remaining contours
Adjust each $B_{ik}$ to remove border offset
Crop patches $\{I_{ik}\}$ from $I$ based on $\{B_{ik}\}$
return $\{I_{i1}, I_{i2}, \ldots, I_{im}\}$
```

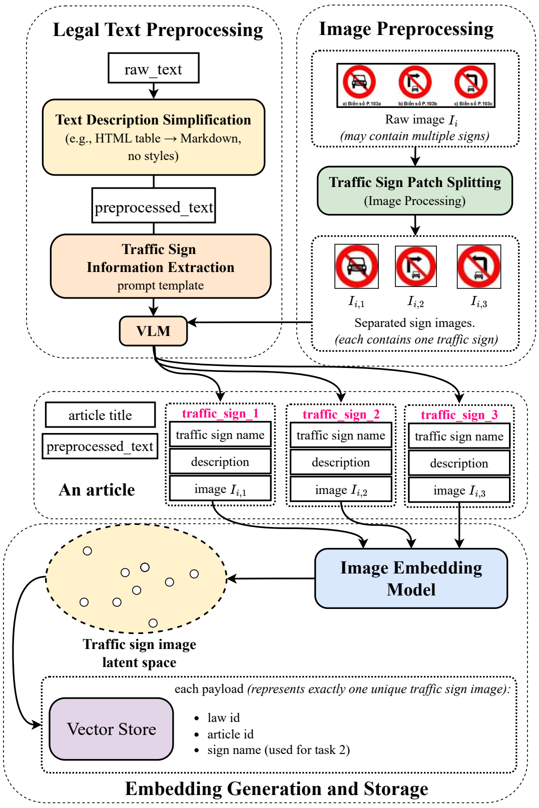


Figure 3: Preprocessing workflow for each article, comprising Legal Text Preprocessing, Image Processing, and Embedding Generation and Storage.

3.2 Approach for Subtask 1: Multimodal Retrieval

Each sample in our Subtask 1 dataset consists of a real-world road scene image I_Q and a corresponding question T_Q.

Traffic sign detection: We first detect all traffic signs in the image using a traffic sign detection model, such as YOLOE (Wang et al., 2025) and GroundingDINO (Liu et al., 2023). Detected signs are cropped based on confidence thresholds and the image resolution ratio.

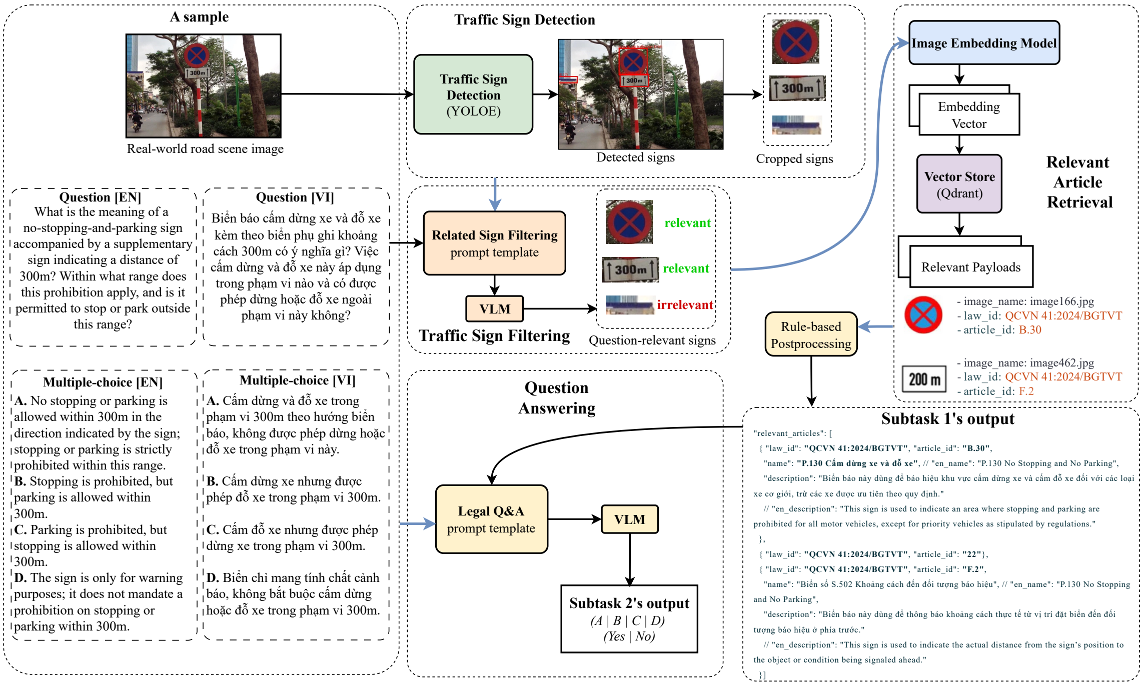
Traffic sign filtering: The cropped traffic signs is embedded within a prompt template together with the question text and the original image. A VLM (e.g., Gemma-3-12B (Team et al., 2025)) is then applied to filter out irrelevant traffic signs, retaining only those pertinent to the given question. For example, as shown in Figure 4, three traffic signs may be detected, but only the first two are relevant; the irrelevant one is discarded.

Relevant article retrieval: The remaining relevant traffic signs are projected into a latent representation space using an image embedding model (e.g., SigLIP2 (Tschannen et al., 2025), OpenAI CLIP (Radford et al., 2021)). Subsequently, for each relevant traffic sign, we retrieve the top-1 article from the vector database (Han et al., 2023) that we constructed in the preprocessing (Section 3.1). For instance, if two relevant signs are retained, this step yields two candidate articles. We then apply a rule-based post-processing step to refine results and remove duplicates, keeping only one instance per article.

3.3 Approach for Subtask 2: Traffic sign rule question answering

Each sample in the Subtask 2 dataset extends upon Subtask 1 by including additional multiple-choice answers. The objective of this subtask is not only to retrieve relevant legal articles but also to guide the VLM in selecting the most appropriate answer option.

The initial article retrieval procedure follows the same algorithm described in the previous subsection, where traffic signs are detected, embedded, and matched with the closest articles in the law database. However, unlike Subtask 1—which focuses solely on retrieving article identifiers—this



Question [EN]
What is the meaning of a no-stopping-and-parking sign accompanied by a supplementary sign indicating a distance of 300m? Within what range does this prohibition apply, and is it permitted to stop or park outside this range?

Question [VI]
Biển báo cấm dừng xe và đỗ xe kèm theo biển phụ ghi khoảng cách 300m có ý nghĩa gì? Việc cấm dừng và đỗ xe này áp dụng trong phạm vi nào và có được phép dừng hoặc đỗ xe ngoài phạm vi này không?

Multiple-choice [EN]
A. No stopping or parking is allowed within 300m in the direction indicated by the sign; stopping or parking is strictly prohibited within this range.
B. Stopping is prohibited, but parking is allowed within 300m.
C. Parking is prohibited, but stopping is allowed within 300m.
D. The sign is only for warning purposes; it does not mandate a prohibition on stopping or parking within 300m.

Multiple-choice [VI]
A. Cấm dừng và đỗ xe trong phạm vi 300m theo hướng biển báo, không được phép dừng hoặc đỗ xe trong phạm vi này.
B. Cấm dừng xe nhưng được phép đỗ xe trong phạm vi 300m.
C. Cấm đỗ xe nhưng được phép dừng xe trong phạm vi 300m.
D. Biển chỉ mang tính chất cảnh báo, không bắt buộc cấm dừng hoặc đỗ xe trong phạm vi 300m.

- image_name: image166.jpg
- law_id: QCVN 41:2024/BGTVT
- article_id: B.30

- image_name: image462.jpg
- law_id: QCVN 41:2024/BGTVT
- article_id: F.2

Subtask 1's output

```
"relevant_articles": [
  { "law_id": "QCVN 41:2024/BGTVT", "article_id": "B.30",
    "name": "P.130 Cấm dừng xe và đỗ xe", // "en_name": "P.130 No Stopping and No Parking",
    "description": "Biển báo này dùng để báo hiệu khu vực cấm dừng xe và cấm đỗ xe đối với các loại xe cơ giới, trừ các xe được ưu tiên theo quy định."
    // "en_description": "This sign is used to indicate an area where stopping and parking are prohibited for all motor vehicles, except for priority vehicles as stipulated by regulations."
  },
  { "law_id": "QCVN 41:2024/BGTVT", "article_id": "22"},
  { "law_id": "QCVN 41:2024/BGTVT", "article_id": "F.2",
    "name": "Biển số S.502 Khoảng cách đến đối tượng báo hiệu", // "en_name": "P.130 No Stopping and No Parking",
    "description": "Biển báo này dùng để thông báo khoảng cách thực tế từ vị trí đặt biển đến đối tượng báo hiệu ở phía trước."
    // "en_description": "This sign is used to indicate the actual distance from the sign's position to the object or condition being signaled ahead."
  }]
```

Figure 4: Our proposed pipeline addresses Subtasks 1 and 2. Subtask 1 focuses on retrieving the most relevant articles, while Subtask 2 leverages the output of Subtask 1 to provide context on Vietnamese traffic signs when the open-source model lacks knowledge of local traffic laws—especially directional information such as left, right, or straight-ahead indicators on traffic signs.

subtask incorporates the traffic sign descriptions into the prompt. The descriptions provide essential context to the VLM, indicating that a real-world road scene image may contain multiple traffic signs along with their corresponding meanings.

Integrating the traffic sign descriptions into the prompt is necessary because the available open-source VLMs lack privileged access to Vietnamese traffic laws and regulations. By embedding descriptive information, we ensure that the model receives sufficient semantic grounding to interpret both the visual content of the image and its legal implications when choosing among the multiple-choice answers.

4 Experiments

4.1 Evaluation Metrics

We evaluate the performance of our approach on two subtasks using task-specific metrics.

4.1.1 Subtask 1: F$_2$ Score

Let $i \in \{1, 2, \ldots, N\}$ denote the index of a sample, where N is the total number of samples. The final F_2 score over the dataset is obtained by averaging across all samples where each sample score $F_{2,i}$ emphasizes recall more heavily than precision:

$$F_2 = \frac{1}{N} \sum_{i=1}^{N} \frac{5 \cdot \text{Precision}_i \cdot \text{Recall}_i}{4 \cdot \text{Precision}_i + \text{Recall}_i}$$

4.1.2 Subtask 2: Accuracy

For multiple-choice prediction, let y_i be the ground-truth label of sample i and $\hat{y}_i$ be the predicted label.

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{y_i = \hat{y}_i\}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function that equals 1 if the condition holds and 0 otherwise.

4.2 Implementation

We conducted experiments for traffic sign detection using the YOLOE (Wang et al., 2025) and GroundingDINO (Liu et al., 2023) models. The detection process was guided by a set of descriptive prompts specifically designed for different traffic sign categories, including: *"blue rectangle traffic sign"*, *"red circular traffic sign"*, *"blue circular traffic sign"*, *"red triangle traffic sign"*, *"green rectangle traffic sign"*, and *"white rectangle text traffic sign"*.

We employed the `Gemma-3-12b-it` model (Team et al., 2025) for traffic sign information extraction, filtering, and question answering through

Detection (LawDB)	Detection (Question)	Embedding Model	Precision	Recall	F2
GroundingDINO	GroundingDINO	SigLIP2	**0.543**	0.546	0.531
GroundingDINO	YOLOE	SigLIP2	0.524	0.573	0.545
GroundingDINO	YOLOE	CLIP	0.542	0.588	0.560
ImageProcessing	YOLOE	CLIP	0.542	**0.596**	**0.566**

Table 2: Experiments for Subtask 1. The first column lists the traffic sign detection models used for LawDB, whereas the second column shows the models used for detecting signs in real-world road scene images.

Base model	Model name	#params	Module
GroundingDINO (Liu et al., 2023)	base	232M	Traffic Sign Detection
YOLOE (Wang et al., 2025)	yoloe-v8l-seg	53M	
SigLIP2 (Wang et al., 2025)	so400m-patch14-384	1.1B	Image Embedding
CLIP (Radford et al., 2021)	CLIP-GmP-ViT-L-14	428M	
Gemma-3 (Team et al., 2025)	gemma-3-12b-it	12.2B	Traffic Sign Filtering and Question Answering

Table 3: All open-source models used in our experiments.

Algorithm 3 Question-Conditioned Retrieval via Sign Detection, VLM Filtering, and Vector Search

```
Require: Road-scene image $I_Q$, question text $T_Q$; detector $\mathcal{D}$; vision–language model $\mathcal{M}$; image embedder $f$; vector DB $\mathcal{V}$ (built per Algorithm 1); post-processing rules $\mathcal{R}$
Ensure: Ranked, de-duplicated set of candidate law articles $\mathcal{A}^\star$
    I. Detect & crop signs
    $B \leftarrow \mathcal{D}(I_Q)$    ▷ Bounding boxes with scores
    $S \leftarrow \{\,\mathrm{crop}(I_Q, b) \mid b \in B,\ \mathrm{score}(b) \geq \tau_{\mathrm{det}},\ \mathrm{passes\_res\_ratio}(b)\,\}$
    II. Filter signs with VLM given the question
    $P \leftarrow \mathrm{PromptTemplate}(S, I_Q, T_Q)$
    $\hat{S} \leftarrow \mathcal{M}.\mathrm{Relevance}(P)$
    III. Embed relevant signs & retrieve candidate articles
    $\mathcal{C} \leftarrow \emptyset$ ▷ Multiset of (article, score) candidates
    for each $I_s \in \hat{S}$ do
        $e \leftarrow f(I_s)$
        $(a^{\mathrm{top}}, \sigma) \leftarrow \mathcal{V}.\mathrm{Top1}(e)$    ▷ Nearest neighbor article & similarity
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{(a^{\mathrm{top}}, \sigma)\}$
    end for
    Rule-based refinement & de-duplication
    $\mathcal{C}' \leftarrow \mathrm{ApplyRules}(\mathcal{C}, \mathcal{R}, I_Q, T_Q)$    ▷ e.g., type constraints, jurisdiction, tie-breaks
    $\mathcal{A}^\star \leftarrow \mathrm{UniqueArticles}(\mathcal{C}')$
    return $\mathcal{A}^\star$
```

the `llama-cpp` server[1]. This model was utilized at multiple stages of the pipeline: (i) preprocessing, (ii) traffic sign filtering for both subtasks, and (iii) question answering, which is specifically required for Subtask 2. To ensure reproducibility, the temperature parameter was fixed at zero.

For image embedding, we experimented with two models: `SigLIP2` (`so400m-patch14-384`) (Tschannen et al., 2025) and `CLIP-GmP-ViT-L-14`[2]. These embeddings were used to capture high-dimensional semantic representations of traffic sign images for downstream tasks. Qdrant[3] is employed as the vector database to ensure efficient and fast data retrieval.

All models were deployed on dual NVIDIA RTX 3060 GPUs, each with 12GB of memory, ensuring efficient inference and reliable reproducibility. The specific model names are summarized in the Table 3.

4.3 Experiment results

After conducting several experiments, we gained several key insights.

First, as shown in Table 2, we found that YOLOE outperforms GroundingDINO on real-world road scene images provided in the questions. The observed increase of 2.6% in the F2 score empirically demonstrates the superiority of YOLOE in this context. *Second*, contrary to ex-

[1] `https://github.com/ggml-org/llama.cpp`

[2] `https://hf.co/zer0int/CLIP-GmP-ViT-L-14`

[3] `https://qdrant.tech`

	F2		Accuracy
OurTeam	**0.6455**	Team1	0.863
TeamA	0.6114	Team2	0.8356
TeamB	0.5992	Team3	0.7808
TeamC	0.579	Team4	0.7329
TeamD	0.5432	Team5	0.726
TeamE	0.4512	Team6	0.7123
TeamF	0.2459	**OurTeam**	**0.6712**
TeamG	0.2385	Team7	0.6233
TeamH	0.1548	Team8	0.6096

Table 4: Leaderboards of the subtasks 1 *(left)* and 2 *(right)*

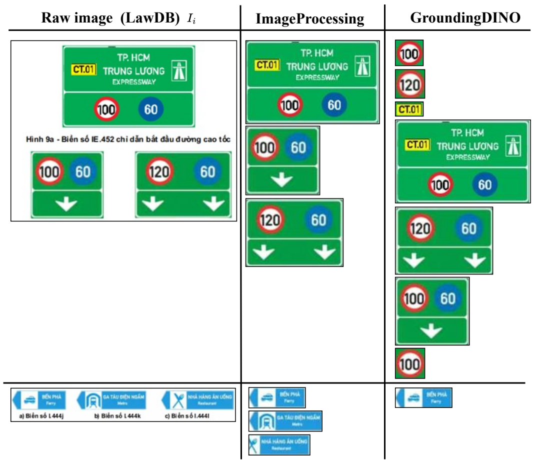


Figure 5: Comparison of Image Processing and GroundingDINO results for a LawDB image with a plain white background. Image Processing demonstrates robust sign image patch splitting compared to GroundingDINO for retrieval purposes.

pectations, the top-performing SigLIP2 model performed slightly worse than `CLIP-GmP-ViT-L-14` in embedding and retrieval tasks by 2.75%. *Lastly*, because LawDB contains isolated traffic sign images with plain white backgrounds, conventional image processing based on rule-based methods outperforms trained models (e.g., GroundingDINO, YOLOE) for extracting and segmenting signs from images. Replacement of GroundingDINO with rule-based image processing significantly increases the final F2 score by 1.1%. Based on these findings, we conclude that the most effective configuration combines rule-based image processing for traffic sign detection in LawDB, YOLOE for traffic sign detection in real-world road scene images, and `CLIP-GmP-ViT-L-14` for embedding traffic sign images. This integrated approach achieved the highest score on the leaderboard for the first subtask.

We further analyze two typical failure types of GroundingDINO on pure LawDB images with plain white backgrounds. As illustrated in Figure 5, GroundingDINO tends to over-partition the sign patch regions (e.g., smaller signs within a larger sign), resulting in redundant or fragmented sign components. Another example involves missing certain traffic signs that are not widely recognized globally, but are iconic in Vietnamese traffic sign notation. Conventional image processing (Algorithm 2), on the other hand, segments traffic signs into distinct and coherent patches that preserve their semantic structure, which is crucial for retrieval.

5 Conclusion

In this work, we introduced **LexiSignVQA**, a unified, training-free, multi-stage framework for multimodal legal question answering on traffic sign rules. Our approach integrates conventional image processing, traffic sign detection, vision–language modeling, and image embedding to address both multimodal retrieval and question answering tasks. Through extensive experiments on the VLSP 2025 MLQA-TSR dataset, our method demonstrated competitive performance, ranking first in Subtask 1 and seventh in Subtask 2.

Beyond quantitative results, our analysis provided valuable insights into the relative strengths of rule-based versus learning-based detection methods and highlighted the challenges of aligning directional information between visual and textual modalities. These findings underscore the importance of combining lightweight, domain-aware preprocessing with modern multimodal learning techniques for robust legal VQA systems.

Future work will focus on improving the handling of directional semantics, enhancing cross-lingual adaptability, and exploring hybrid training strategies to further advance the interpretability and reliability of multimodal systems in legal and safety-critical contexts.

References

Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In Conference and Labs of the Evaluation Forum.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. 2015. Vqa: Visual question answering. International Journal of Computer Vision, 123:4 – 31.

Amer Farea and Frank Emmert-Streib. 2025. Understanding question-answering systems: Evolution, applications, trends, and challenges. Eng. Appl. Artif. Intell., 156:110997.

Yikun Han, Chunjiang Liu, and Pengfei Wang. 2023. A comprehensive survey on vector database: Storage and retrieval technique, challenge. arXiv preprint arXiv:2310.11703.

A. Jabri, Armand Joulin, and Laurens van der Maaten. 2022. Visual question answering: From theory to application. Visual Question Answering.

Hoa Quang Le, Huong Xuan Dieu Kieu, Khiem Vinh Tran, and Binh Thanh Nguyen. 2024. Lawvivqa: A visual question answering dataset for vietnamese legal content. 2024 RIVF International Conference on Computing and Communication Technologies (RIVF), pages 393–397.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chun yue Li, Jianwei Yang, Hang Su, Jun-Juan Zhu, and Lei Zhang. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European Conference on Computer Vision.

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. In International Conference on Machine Learning.

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.

The Association for Vietnamese Language and Speech Processing VLSP. 2025. VLSP 2025 Challenge on Multimodal Legal QA on Traffic Sign Rules.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. ArXiv, abs/2502.14786.

Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Yoloe: Real-time seeing anything. ArXiv, abs/2503.07465.

A Appendix

A.1 Traffic Sign Information Extraction prompt template

In this section, we formulate the prompt template to guide a vision-language model (VLM) in parsing traffic law articles and extracting structured information about traffic signs. The structure of the prompt is shown in Figure 6.

- **Contextual Integration**: The prompt incorporates both the article title and the full textual content, thereby ensuring that the model receives sufficient contextual information to accurately interpret and extract traffic sign data.
- **Schema-Constrained Output**: The use of a predefined JSON schema enforces structural consistency across outputs, which facilitates reliable downstream parsing and post-processing.
- **Index-Referenced Alignment**: By explicitly specifying the start and end indices of the traffic sign images, the framework guarantees completeness of extraction while preventing duplication or omission.
- **Language Specification**: The prompt mandates that both the name and description fields be produced exclusively in Vietnamese, thereby aligning the output with the linguistic domain of the source regulation.
- **Controlled Conciseness**: The model is instructed to avoid any auxiliary commentary, returning only the structured JSON array, which minimizes verbosity and enhances reproducibility.

A.2 Related Sign Filtering prompt template

This section describes the prompt template used for answering legal questions concerning traffic and road regulations while reasoning over visual evidence from detected traffic signs. The structure of the prompt is shown in Figure 7.

- **Positional Reference Extraction**: The prompt explicitly instructs the model to determine whether the question refers to signs at particular spatial positions (e.g., left/right, top/bottom, near/far). If no such reference is present, the model must return "No explicit

SIGN NAME PARSING (LAWDB PROCESSING)

```
You are given the title and content of an article from the Vietnam National Technical Regulation on Traffic Signs and Signals.

TITLE: <<TITLE>>

CONTENT:
<<CONTENT>>

Your task is to extract and structure information about all traffic signs (<<IMAGE_<int>>>) mentioned in the article. Output the
result strictly in JSON format as an array of objects, following the schema below:

[
   {
      "image_tag": "<<IMAGE_0>>" # image identifier
      "name": "<The name of the traffic sign in Vietnamese>",
      "description": "<Description in Vietnamese: What does this sign indicate or warn about?>",
   },
   {
      "image_tag": "<<IMAGE_1>>" # image identifier
      "name": "<The name of the traffic sign in Vietnamese>",
      "description": "<Description in Vietnamese: What does this sign indicate or warn about?>",
   },
   …
]

Requirements:
- The traffic signs are in the order extracted from the article.
- Each object represents one unique traffic sign. There must be <<NUM_SIGNS>> objects in the resulting array from
<<IMAGE_<FROM_INDEX>>> to <<IMAGE_<TO_INDEX>>>.
- Images in the content are defined as <<IMAGE_<index>>> (e.g., <<IMAGE_0>>, <<IMAGE_1>>, etc.)
- Use only Vietnamese for the name and description fields.
- The image field should be in the format "<<IMAGE_<index>>>".
- Use \" instead of " inside double quotes. Do not include any explanatory text—only return the resulting JSON array.
```

Figure 6: Traffic Sign Information Extraction prompt template

position" thereby ensuring clarity in positional grounding.

- **Visual Feature Identification**: The model is required to extract descriptive attributes of signs mentioned in the question, such as color, shape, symbols, or icons. Logical conjunction (AND) is enforced when multiple attributes are specified, which guarantees precise matching. A strict color interpretation rule is also incorporated (e.g., "màu xanh" is consistently mapped to "blue").

- **Comprehensive Scene Description**: The prompt mandates that the model generate a concise yet complete description of the entire image, covering the road layout, environment, vehicles, pedestrians, and the relative positions of all detected signs. This step ensures contextual completeness before individual sign evaluation.

- **Sign-wise Reasoning and Justification**: For each detected sign, the model must assess relevance to the posed question, integrating both spatial alignment and semantic features. Supplementary rectangular signs containing text are considered when determining the scope of prohibitory signs, preventing misinterpretation.

- **Boolean-List Decision Schema**: The final decision is expressed as a Python-style boolean list, with one entry per detected sign, in the order they are provided. The framework enforces non-triviality by requiring at least one True value; in the absence of a clear match, the most salient sign is marked True.

- **Output Format Control**: The model is constrained to produce results in a strictly defined textual format—beginning with structured reasoning components (positions, visual features, explanations), followed by the final boolean list enclosed between «ANSWER» and «/ANSWER». This format prevents verbosity while ensuring reproducibility and downstream parsability.

A.3 Legal Q&A prompt template

We introduce the prompt template used for answering legal questions for Subtask 2. The structure of the prompt is shown in Figure 8.

EXTRACTED SIGNS FILTERING (REMOVE IRRELEVANT DETECTED SIGNS FROM INPUT IMAGE)

You are an expert in legal question answering, specializing in traffic and road-related regulations.

You will be provided with:
1. A question related to traffic or road regulations.
2. An original input image.
3. <<NUM_SIGNS>> detected sign(s) from the given image.

Your task:
1. Read the question carefully and determine whether it refers to a sign at a specific position. Explicitly extract the referenced position(s) if present (e.g., left/right/top/bottom/center; near/far; overhead/ahead/behind). If none, state "No explicit position".
2. From the question, identify the expected visual characteristics of the referenced sign, including (if mentioned or implied) its shape, color(s), symbol(s), icon(s), or other notable features. If none, state "No explicit visual features". IMPORTANT: If the question mentions multiple descriptive features (e.g., color AND shape AND icon), then ALL of them must be satisfied simultaneously (logical AND), NOT just one (NOT logical OR). IMPORTANT COLOR RULE: In Vietnamese, the phrase "màu xanh" MUST be interpreted as "blue". Example: If the question says "biển màu xanh", a blue-colored sign will satisfy this condition.
3. Give a concise but complete description of the entire original image, noting road layout, vehicles, pedestrians, environment, and the positions of all detected signs (left/right/top/bottom/center).
4. For EACH detected sign (in the given order), decide whether it is related to the question.
5. When deciding, pay close attention to:
+ The position of the sign in the original image relative to the viewpoint and to any position(s) referenced by the question.
+ Whether the question explicitly or implicitly refers to a specific location or direction (e.g., “sign on the right”, “overhead sign”).
+ The visual appearance and meaning of the sign.
+ If the question is about a prohibitory sign, also consider any supplementary sign(s) immediately below it that is in rectangle shape, contain text or, as they may modify the prohibition’s scope.
6. For EACH sign, explain your reasoning clearly and briefly, including position relevance if applicable.
7. Provide the FINAL decision as a Python-style list of boolean values (True/False), where:
+ EACH ELEMENT MUST CORRESPOND EXACTLY to the matching detected sign in the SAME ORDER they were provided.
+ The length of the list MUST equal <<NUM_SIGNS>>.
8. STRICT REQUIREMENT: The final boolean list MUST contain at least one True value. If no detected sign clearly matches the question, then choose the single most prominent/main sign in the image (e.g., the largest or most central sign) and mark it as True.
9. Enclose ONLY the final boolean list between <<ANSWER>> and <</ANSWER>> tags, with nothing else inside.
Output format:
Question-referenced position(s): …
Question-referenced visual features: …
Full image description: …
Explanation for sign 1: …
Explanation for sign 2: …
…

```
<<ANSWER>[
   <True/False answer for sign 1>,
    …
   <True/False answer for sign <<NUM_SIGNS>>>
]<</ANSWER>>
```

IMPORTANT NOTES:
- 4-wheeled vehicles may include car, truck, van, bus, jeep, …
- 3-wheeled vehicles may include tricycle, auto-rickshaw, cycle rickshaw, …
- The question may ask about multiple signs, not only one sign.
- "màu xanh" = blue in Vietnamese
- "phương tiện"/"loại xe" = all pedestrians, bicycles, cars, trucks, motorbikes, auto-rickshaws, ...

Figure 7: Related Sign Filtering prompt template

Legal Q&A prompt template

MULTIPLE CHOICE QUESTION	YES / NO QUESTION
You are a Legal QA Assistant. You will be given question, image and multiple choices and must to choose 1 answer. Explain the answer before return the final selection (A, B, C or D) inside <answer> and </answer>.	You are a Legal QA Assistant. You will be given a yes/no question, image must to answer. Explain the answer before return the final selection (Yes or No) inside <answer> and </answer>.

Figure 8: Legal Q&A prompt template