Metamorphic at VLSP 2025: SIGMA – A Multimodal Agent System for Legal QA on Vietnamese Traffic Signs

Nguyen Tuan Kiet[1,2]* **and Nguyen Khanh Tuan Anh**[1,2] **and Long Hoang Huu Nguyen**[1,2]
Dam Vu Trong Tai[1,2] and **Dang Van Thin**[1,2]
[1]University of Information Technology-VNUHCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City, Vietnam
{21521042, 22520055, 22520817}@gm.uit.edu.vn,
{taidvt, thindv}@uit.edu.vn

Abstract

We present SIGMA (SIGn Multimodal Agents), a system developed for the VLSP 2025 MLQA-TSR shared task on multimodal legal question answering with Vietnamese traffic sign rules. The task requires integrating traffic sign images and legal documents to answer multiple-choice and yes/no questions, demanding both precise visual interpretation and reliable legal grounding. Our approach adopts a multi-agent architecture that combines traffic sign understanding, legal text retrieval, and reasoning over both modalities. On the official evaluation, SIGMA achieved 72% accuracy in Subtask 2, ranking among the top-5 finalists. These results demonstrate the effectiveness of agent-based multimodal approaches for answering legal questions in safety-critical domains such as traffic law compliance.

1 Introduction

Visual Question Answering (VQA) is a multimodal task that combines computer vision (CV) and natural language processing (NLP), where systems are required to answer natural language questions about an image. In the most common form of VQA, the model is presented with an image and a textual question and must then determine the correct answer, typically expressed as a few words or a short phrase. Unlike traditional CV tasks such as classification or detection, VQA is inherently open-ended: the form and semantics of the question are not fixed in advance, demanding both visual understanding and language reasoning. Variants of the task include binary (Yes/No) questions (Antol et al., 2015; Zhang et al., 2016) and multiple-choice formats (Antol et al., 2015; Zhu et al., 2016), in which candidate answers are proposed to the system.

*Corresponding author. Email: 21521042@gm.uit.edu.vn

Building on this paradigm, the VLSP 2025 – MLQA-TSR (Multimodal Legal Question Answering on Traffic Signs and Regulations) shared task extends VQA to the legal domain, focusing on Vietnamese traffic sign interpretation and regulations. The task consists of two subtasks:

- **Multimodal Retrieval**, which requires retrieving relevant legal articles given a traffic sign image and a natural language question
- **Question Answering**, where the system must answer multiple-choice (A, B, C, D) or Yes/No questions about traffic signs and scenarios.

In this work, we concentrate on **Subtask 2: Question Answering**, and propose an unsupervised multimodal *multi-agent* approach. Instead of task-specific fine-tuning, our system leverages specialized agents that collaborate to interpret traffic signs, integrate legal knowledge, and perform multimodal reasoning to answer questions accurately.

The main contributions of this paper are as follows:

- We propose **SIGMA**, a multimodal multi-agent system designed for legal VQA on Vietnamese traffic sign regulations, addressing Subtask 2 of the VLSP 2025 – MLQA-TSR shared task.
- Our approach adopts an **unsupervised setting**, avoiding task-specific supervised fine-tuning while leveraging modular agents for sign understanding, legal text retrieval, and multimodal reasoning.
- We demonstrate that a multi-agent architecture can achieve competitive performance in a safety-critical application, obtaining 72% accuracy and ranking in the top-5 systems in the official evaluation.

The rest of this paper is organized as follows. Section 2 reviews related work on VQA, multimodal QA, and legal reasoning. Section 3 describes our proposed multi-agent methodology. Section 4 outlines the experimental setup, while Section 5 reports the main results and discussions. Finally, Section 6 concludes the paper and discusses potential future directions.

2 Related Work

Early research in Visual Question Answering (VQA) primarily combined convolutional neural networks (CNNs) for image representation with recurrent neural networks (RNNs) for question encoding, learning a joint embedding space for answer prediction (Gao et al., 2015; Malinowski et al., 2015; Ma et al., 2016). Later works introduced attention mechanisms to align specific image regions with relevant parts of the question (Zhu et al., 2016; Xu and Saenko, 2016; Chen et al., 2015; Jiang et al., 2015; Andreas et al., 2016; Yang et al., 2016), significantly improving performance. Several VQA variants have been studied, including binary yes/no questions (Antol et al., 2015; Zhang et al., 2016) and multiple-choice settings (Antol et al., 2015; Zhu et al., 2016), alongside open-ended formulations. Benchmarks such as DAQUAR (Malinowski and Fritz, 2014), COCO-QA (Ren et al., 2015), The VQA Dataset (Antol et al., 2015), FM-IQA (Gao et al., 2015), Visual7W (Zhu et al., 2016), and Visual Genome (Krishna et al., 2017) have driven progress in this field by providing diverse question formats and evaluation settings.

Beyond traditional VQA, multimodal question answering has expanded toward integrating external knowledge and reasoning capabilities. Retrieval-augmented approaches combine visual understanding with textual grounding to answer knowledge-intensive questions. More recently, large vision-language models such as BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2024), and Gemini (Team et al., 2025a) demonstrate strong zero-shot performance on multimodal reasoning tasks, showing the potential of pretrained architectures for generalization. In parallel, modular and multi-agent systems have been explored to decompose complex tasks into specialized subtasks, providing interpretability and flexibility for multimodal reasoning pipelines.

While prior research in VQA has developed strong supervised methods, and traffic sign recognition has matured as a vision problem, little work has addressed traffic sign reasoning in a multimodal QA setting. Existing approaches are often dataset-specific and require extensive labeled data, limiting generalization. Moreover, multi-agent or unsupervised frameworks have not been systematically applied to this task. Our work seeks to fill this gap by proposing an unsupervised multimodal multi-agent system tailored for traffic sign QA, providing a novel direction for scalable and interpretable solutions.

3 Methodology

We propose **SIGMA** (SIGn Multimodal Agents), a modular multi-agent system for Subtask 2 (Question Answering) of VLSP 2025 – MLQA-TSR. The system decomposes multimodal legal QA into two cooperating subgraphs—*Image Subgraph* and *Article Subgraph*—whose outputs are fused by a *Reasoning Agent* to produce the final answer with legal citations. An overview is shown in Figure 1.

3.1 Problem Formulation

Given an input image $\mathcal{I}$ containing one or more traffic signs and a natural-language question q (either multiple-choice with options $\mathcal{C} = \{A, B, C, D\}$ or binary Yes/No), the system must output an answer $y \in \mathcal{C}$ or $y \in \{\text{Yes}, \text{No}\}$, together with short supporting citations from Vietnamese traffic regulations: the *Law on Road Traffic Order and Safety (36/2024/QH15)* and the *National Technical Regulation on Traffic Signs and Signals (QCVN 41:2024/BGTVT)*. The core challenge is to combine precise visual understanding of traffic signs with legally grounded reasoning over textual regulations.

3.2 System Overview

SIGMA follows a perception–retrieval–reasoning pipeline. The **Image Subgraph** detects signs and converts visual content into structured attributes and a scene-level description. The **Article Subgraph** retrieves candidate legal passages and filters them down to the minimal set of relevant clauses. The **Reasoning Agent** integrates both sources to select an answer and emit supporting citations. All components are orchestrated as graph nodes (agents) with explicit edges for data flow; Section 3.7 details the orchestration.

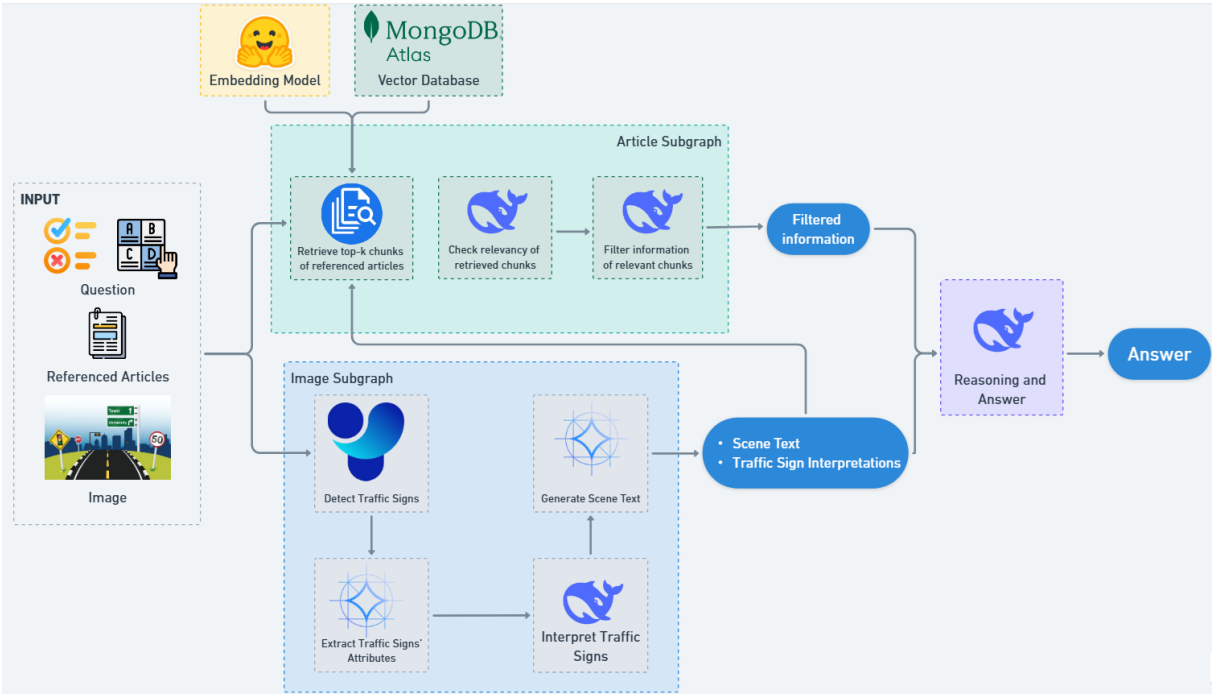

Figure 1: System overview of **SIGMA**.

3.3 Image Subgraph

(i) Sign Detection. We first run a YOLO-based detector on the input image $\mathcal{I}$ to localize traffic signs. The detector returns bounding boxes $\{b_i\}$ and coarse sign categories (e.g., regulatory, warning, direction). Crops from $\{b_i\}$ are passed downstream.

(ii) Attribute Extraction & Scene Text. Each detected crop is processed by a vision-language model to extract fine-grained *attributes* such as *shape* (round, triangular), *color scheme* (red border, blue background), *pictogram* (left-turn arrow, pedestrian), and *numerical tokens* (e.g., "50"). In parallel, the VLM generates a concise *scene description* for $\mathcal{I}$ that summarizes salient traffic elements (e.g., "urban road, blue circular sign with 50 indicating minimum speed"). Outputs are serialized as a compact schema:

```
{
  "sign_id": "...",
  "shape": "..."
  "background_color": "...",
  "border": "...",
  "icon": "...",
  "text": "...",
  "diagonal_line": "...",
  "sub_sign": "..."
}
```

Output schema of sign attributes

(iii) Sign Interpretation. To bridge visual attributes and legal semantics, a lightweight LLM interprets each sign into a canonical textual meaning (e.g., "No entry", "Minimum speed 50 km/h", "Pedestrian crossing") and, when applicable, enumerates conditions/modifiers (e.g., "except buses", "in effect 7–9 AM"). The interpreter consumes the attribute schema + scene text and outputs:

```
{
  "type": "...",
  "meaning": "...",
},
```

Output schema of scene text

(iv) Scene Description. Finally, the VLM generates a concise *scene-level description* conditioned on the set of interpreted signs, yielding context such as: *"Urban intersection with a circular blue sign indicating minimum speed 50 and a supplementary panel restricting motorcycles."*

This language-first representation enables downstream retrieval and reasoning without requiring image features.

3.4 Article Subgraph

(i) Corpus Ingestion & Vector Store. We index two corpora: *36/2024/QH15* and *QCVN 41:2024/BGTVT*. Documents are split into overlapping chunks (by article/section; optionally with a token window). Each chunk is embedded using the law-domain-tuned model and stored in a MongoDB Atlas *vector database* with metadata (law id, article id, title, text).

(ii) Retrieval of Candidate Chunks. At query time, we concatenate the user question q with image-side text (*scene-level description* and *canonical sign meanings*). The combined text is embedded by the same embedding model and used to retrieve top-k candidate chunks from MongoDB Atlas via vector similarity. This yields a ranked list $\mathcal{R} = \{(c_j, s_j)\}_{j=1}^{k}$ of legal text chunks c_j with similarity scores s_j.

(iii) Relevancy Filtering. Since retrieval can include noise, we employ an LLM to filter $\mathcal{R}$ and return only the most relevant legal passages for answering q. The model discards off-topic or redundant chunks and outputs a reduced set $\mathcal{L} = \{c_{j'}\}$, preserving only legally grounded snippets. In this work, we use $\mathcal{L}$ directly as the input to the Reasoning Agent, without further semantic compression.

3.5 Reasoning and Answering

The **Reasoning Agent** receives (i) sign interpretations from the Image Subgraph and (ii) filtered legal chunks $\mathcal{L}$ from the Article Subgraph. Based on question type, the agent produces:

Output schema of the Reasoning Agent

```
# Multiple-choice format
{
  "answer": "A" | "B" | "C" | "D"
}

# Yes/No format
{
  "answer": "Yes" | "No"
}
```

The prompt explicitly instructs the agent to ground its prediction in $\mathcal{L}$ and avoid uncontrolled world knowledge. When $\mathcal{L}$ lacks sufficient evidence, the agent is encouraged to abstain or return "uncertain". In the shared-task evaluation, unresolved uncertainty defaults to the highest-ranked candidate from $\mathcal{L}$.

3.6 Input/Output Contracts & Data Flow

Subgraph Contracts. We formalize each subgraph as a mapping function:

$$\begin{aligned} f_{\text{img}} &: \mathcal{I} \longrightarrow (\, s,\ \{m_i\}_{i=1}^{k}\,) \\ f_{\text{art}} &: (q, s, \{m_i\}_{i=1}^{k}) \longrightarrow \mathcal{L} \\ f_{\text{reason}} &: (s, \{m_i\}_{i=1}^{k}, \mathcal{L}, \mathcal{C}?) \longrightarrow y \end{aligned} \tag{1}$$

where $\mathcal{I}$ denotes the input image, q the question, s the scene-level description, $\{m_i\}_{i=1}^{k}$ the interpreted meanings of detected signs, $\mathcal{L}$ the retrieved legal chunks, $\mathcal{C}$ the (optional) candidate answers, and y the final predicted answer.

Core Models. The proposed system relies on the following core components:

- **Sign Detector.** A YOLO-based model is employed to localize traffic signs in the input image, providing bounding boxes and coarse categories.
- **Vision–Language Model.** Gemma 3 27B (Team et al., 2025b) is prompted to extract structured attributes (shape, color scheme, pictogram, text, etc.) from detected sign crops and to generate a compact scene-level description. Outputs follow a minimal JSON-like schema with fallbacks (e.g., *digits:null* if absent).
- **Large Language Model.** DeepSeek-R1-0528 (DeepSeek-AI et al., 2025) serves three roles: (i) interpreting extracted attributes into canonical sign meanings, (ii) filtering and validating retrieved legal passages, and (iii) performing the final reasoning step. Prompts enforce structured outputs and discourage unsupported claims.
- **Embedding Model & Vector Store.** Legal text chunks are embedded with the domain-specific model *tranguyen/halong_embedding-legal-document-finetune* and indexed in MongoDB Atlas as a vector database. Metadata (law id, article id, title, text) is preserved to support precise citation.

3.7 Orchestration with LangGraph

SIGMA is implemented with **LangGraph**, which represents the workflow as a *graph* of nodes and edges. In this abstraction, each *node* is an agent (e.g., YOLO detector, VLM parser, legal retriever,

	# Images	# Questions	Yes/No Qs	MCQs
Train	304	530	154	376
Public Test	90	50	18	32
Private Test	104	146	72	74
Total	498	726	244	482

Table 1: Dataset statistics for VLSP 2025 - MLQA - TSR.

Split	# Images	# Instances
Train	209	774
Validation	59	274
Test	30	88
Total	298	1,051

Table 2: Dataset statistics for YOLO fine-tuning.

Category	# Instances
Auxiliary	183
Information	141
Information Expressway	116
Prohibition	468
Regulatory	153
Warning	75
Total	1,051

Table 3: Category-wise distribution of YOLO fine-tuning dataset.

legal filter, reasoner), and an *edge* encodes data flow and execution order. The Image Subgraph and Article Subgraph are defined as reusable *subgraphs* that can run independently and then feed into the global graph. LangGraph maintains a shared *state* object (question, scene text, sign list, candidate chunks, filtered snippets, answer) that nodes read/write, enabling clear contracts and deterministic composition. This design simplifies error isolation, facilitates iterative development of each agent, and makes the overall system extensible (e.g., swapping the detector or embedding model without changing the global control flow).

4 Experimental Setup

YOLO Detector Training. For traffic sign detection, we employ the **Ultralytics**[1] implementation of YOLO. A dataset of traffic scene images was manually labeled with bounding boxes and sign categories. The labeled data was split into training and validation subsets following an 80/20 ratio. We trained YOLO with standard augmentation (random scaling, flipping, and brightness adjustments) and optimized using the default settings of Ultralytics (SGD with momentum, cosine learning rate schedule). The trained detector provides bounding boxes b_i and coarse categories for each sign in an input image.

Vision–Language Model. The vision–language model **Gemma 3 27B** is not fine-tuned; instead, we access it in an inference-only manner through the *FPT AI Marketplace*[2] cloud platform. The model is prompted with structured instructions to extract attributes (shape, color, icon, text, modifiers) and generate a compact scene description. Since no supervised fine-tuning is applied, this stage operates in a zero-shot fashion.

Large Language Model. For both sign interpretation and legal reasoning, we use **DeepSeek-R1** hosted on *Azure AI Foundry*[3] . Similar to Gemma 3, DeepSeek-R1 is used exclusively in inference mode. The model is prompted to (i) filter irrelevant legal passages based on their semantic and syntactic relevance to the extracted traffic sign attributes, and (ii) generate structured answers following a chain-of-thought style reasoning prompt. The exact prompting templates are provided in the Appendix. Importantly, no fine-tuning or additional training is performed—this step is entirely zero-shot prompting.

Embedding Model. All legal documents are semantically chunked and embedded using the domain-specific model tranguyen/halong_embedding-legal-document-finetune, which is derived from E5 (Wang et al., 2024) and available on HuggingFace. This model is likewise used in an inference-only setting, without fine-tuning in our work. Embeddings are indexed in MongoDB Atlas for retrieval.

[1] https://www.ultralytics.com/

[2] https://marketplace.fptcloud.com/

[3] https://ai.azure.com/

Rank	Accuracy
Top 1	0.8630
Top 2	0.8356
Top 3	0.7808
Top 4	0.7328
Top 5 (Ours)	0.7260

Table 4: Performance of top 5 systems on the **private test set**. Our system **Top 5** achieve an accuracy of 0.7260

Supervision Level. Apart from the YOLO detector, which requires supervised training on annotated images, the remainder of the system operates in an **unsupervised** or zero-shot manner. Specifically, the VLM (Gemma 3 27B), LLM (DeepSeek-R1), and embedding model are only guided via carefully engineered prompts. The criteria for filtering irrelevant legal texts and for structuring reasoning outputs are embedded in these prompts (see Appendix), not learned from additional training. This makes SIGMA largely adaptable to new domains without requiring further supervision.

5 Results and Discussion

5.1 Datasets

We evaluate SIGMA on the official shared-task benchmark for traffic sign question answering (VLSP 2025 – MLQA-TSR). Each example consists of an input *traffic scene image*, its associated *referenced legal article*, a natural language *question*, and the *gold answer*. Questions are either *binary (Yes/No)* or *multiple-choice*. The training split is used exclusively to supervise the YOLO detector; all other components (Gemma VLM, DeepSeek-R1 LLM, embedding model) operate in a zero-shot or unsupervised fashion without fine-tuning. The *public test set* is treated as a development set for error analysis, while the *private test set* provided by the organizers is reserved for final evaluation.

The overall dataset composition, including the number of images and question types per split, is summarized in Table 1.

YOLO Fine-tuning Dataset. The detector was fine-tuned on a dataset derived directly from the training split of this shared-task images. From the 304 training images, we manually annotated traffic signs with bounding boxes and categories, resulting in 298 usable images after preprocessing (auto-orientation, resizing to 640×640). The annotated dataset contains 1,051 bounding box instances across 6 traffic sign categories. We split the annotated data into train/validation/test subsets in a 70/20/10 ratio, without applying additional augmentation. Table 2 summarizes the statistics.

5.2 Evaluation metrics

We report results using **accuracy**, defined as the proportion of correctly predicted answers over the total number of questions. Formally, let y_i denote the gold label and $\hat{y}_i$ the predicted label for the i-th question. Accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[y_i = \hat{y}_i] \tag{2}$$

where N is the total number of questions, and $\mathbf{1}\{\cdot\}$ is the indicator function that equals 1 if the prediction matches the gold label and 0 otherwise.

5.3 Results

The performance of our system in the VLSP 2025 – MLQA – TSR shared task, together with the top four teams, is reported in Table 4. On the private test set, our approach achieved an accuracy of 0.7260, placing us 5th overall in the final leaderboard.

Beyond the leaderboard results, this shared task demonstrates the feasibility of our approach: we successfully developed a multimodal, multi-agent system that integrates perception, retrieval, and reasoning to address a challenging visual question answering task in the legal domain. This validates both the design of our pipeline and the effectiveness of combining vision-language models, legal retrieval, and structured reasoning agents for real-world, law-grounded QA.

5.4 Error Analysis

To better understand the limitations of SIGMA, we qualitatively some examined failure cases. We identified three recurring sources of error:

Figure 2: Failure case caused by incorrect **sign type classification**: the detector identified the sign but the sign interpreter mislabeled its category, leading to an incorrect downstream interpretation.

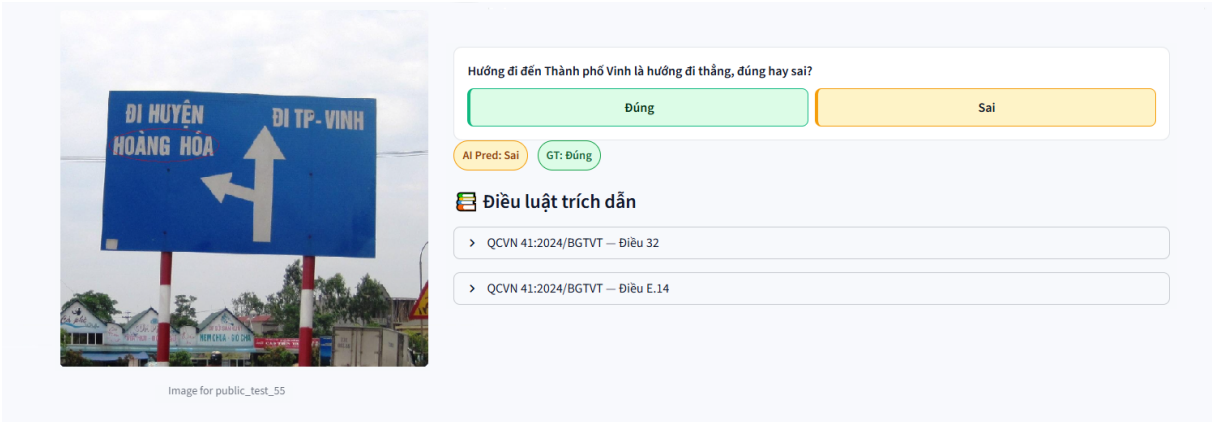

Figure 3: Failure case caused by **ambiguous sign meaning**: the sign interpreter produced wrong sign semantic interpretation, which propagated into the reasoning step.

(i) Sign type misclassification. As illustrated in Figure 2, the detector successfully localized the sign but the interpreter mislabeled its category (e.g., mistaking a regulatory sign for a warning sign). Such errors highlight the sensitivity of the reasoning pipeline to early misclassifications, since all downstream reasoning depends on the canonical meaning.

(ii) Ambiguous sign semantics. In some cases, shown in Figure 3, the interpreter produced a plausible but incorrect semantic meaning (e.g., interpreting a "minimum speed" sign as "maximum speed"). This demonstrates the difficulty of fine-grained semantic distinctions when multiple signs share similar visual patterns.

(iii) Scene context misinterpretation. Figure 4 presents cases where individual signs were correctly recognized, but their meaning was incorrectly applied in the broader scene context (e.g., overlooking supplementary panels or lane restrictions). Such errors indicate that beyond sign recognition, accurate scene understanding remains an open challenge.

Overall, these analyses suggest that future improvements should focus on (a) enforcing stricter constraints between visual attributes and legal semantics, (b) leveraging joint reasoning over multiple signs and contextual cues, and (c) incorporating explicit legal knowledge to disambiguate visually similar categories.

6 Conclusion

We introduced **SIGMA**, a perception–retrieval–reasoning framework for traffic sign question answering grounded in Vietnamese traffic law. The system integrates a supervised detector with unsupervised large-scale models for sign interpretation, legal retrieval, and reasoning. Experimental results demonstrate strong performance without fine-tuning of the language components, underscoring the viability of zero-shot legal reasoning with multimodal inputs.

Future work will focus on (i) expanding coverage to additional traffic sign categories, (ii) improving

Figure 4: Failure case caused by **scene context misinterpretation**: although the sign was correctly recognized, contextual elements (e.g., lane markings, vehicle placement) were misinterpreted, resulting in the wrong answer.

detection robustness under challenging visual conditions, and (iii) integrating chain-of-thought verification to mitigate reasoning errors. Beyond traffic law, the proposed approach offers a general template for multimodal legal QA systems that combine perception, retrieval, and reasoning in a unified pipeline.

References

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia. 2015. Abc-cnn: An attention based convolutional neural network for visual question answering. *arXiv preprint arXiv:1511.05960*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Aiwen Jiang, Fang Wang, Fatih Porikli, and Yi Li. 2015. Compositional memory for visual question answering. *arXiv preprint arXiv:1511.05676*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Lin Ma, Zhengdong Lu, and Hang Li. 2016. Learning to answer questions from image using convolutional neural network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems*, 28.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie

Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025a. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European conference on computer vision*, pages 451–466. Springer.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5014–5022.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A Prompt Templates

Check Article Relevancy Prompt

Bạn là một trợ lý pháp lý giao thông. Dựa trên thông tin sau, hãy đánh giá xem điều luật dưới đây có liên quan đến câu hỏi trắc nghiệm, các lựa chọn hay các phân tích về ảnh hay không.
Câu hỏi: question
Các lựa chọn trả lời:
{choices}
Thông tin nhận diện từ biển báo giao thông:
{sign_interpretation}
Ngữ cảnh trong hình ảnh:
{scene_text}
Nội dung điều luật:
{article_text}
Yêu cầu:
Trả lời dưới dạng JSON đúng định dạng sau, không viết thêm bất kỳ dòng nào khác ngoài JSON:
{
"reason": "<giải thích vì sao điều luật này có hoặc không liên quan đến câu hỏi>",
"relevant": True hoặc False
}
Gợi ý:
- Nếu điều luật giúp loại trừ hoặc lựa chọn được đáp án đúng → relevant = True.
- Nếu điều luật không có thông tin nào giúp trả lời câu hỏi → relevant = False.

Article Filter Information Prompt

Bạn là một trợ lý pháp lý giao thông Việt Nam.
Dựa trên thông tin dưới đây, hãy đọc, phân tích và **tóm tắt** nội dung chính của điều luật, sao cho phù hợp với ngữ cảnh câu hỏi, lựa chọn trả lời và thông tin phân tích biển báo.
Câu hỏi:
{question}
Các lựa chọn trả lời:
{choices}
Thông tin nhận diện từ biển báo giao thông:
{sign$_interpretation$}
Ngữ cảnh trong hình ảnh:
{scene$_text$}
Nội dung điều luật:
{article$_text$}
Yêu cầu:
- Nếu điều luật không liên quan tới ngữ cảnh, vẫn tóm tắt nội dung chính của nó.
- Giữ nguyên ý chính của điều luật nhưng có thể rút gọn, bỏ các chi tiết không cần thiết.
- Chỉ trả lời **duy nhất** bằng JSON thuần, có thể phân tích bằng ‘json.loads()‘.
- Đảm bảo định dạng JSON thuần, không thêm văn bản, giải thích hoặc ký tự thừa như sau:
{
"text": "nội dung tóm tắt ngắn gọn, súc tích, tập trung vào điểm chính liên quan đến ngữ cảnh trên"
}

Sign Attributes Extraction Prompt

Bạn là một chuyên gia phân tích hình ảnh giao thông. Phân tích đặc điểm của biển báo giao thông trong ảnh dưới đây. Hãy mô tả chi tiết các đặc điểm có thể giúp phân biệt và truy xuất thông tin từ một tập biển báo có sẵn:
1. Hình dạng tổng thể của biển báo là gì?
2. Màu nền chính của biển là gì? Có viền không? Nếu có, màu viền là gì?
3. Biển có chứa hình ảnh biểu tượng gì không? (ví dụ: mũi tên, xe, người, v.v.)
4. Biển có chứa chữ hay số không? Nếu có, hãy ghi rõ nội dung chữ/số đó.
5. Biển có gạch chéo đỏ hay không?
6. Có biển phụ kèm theo không? Nếu có, hãy mô tả nội dung và hình dạng của biển phụ.
Yêu cầu: - Chỉ trả lời dưới dạng **JSON**, không kèm theo giải thích hoặc văn bản dư thừa.
- Mỗi trường cần điền càng rõ càng tốt, nếu không thấy thì để ‘null‘.
- Đảm bảo đúng định dạng JSON như ví dụ bên dưới.
Cần trả lời theo cấu trúc JSON như sau: [{
"id": "Biển số của biển báo hoặc null nếu không rõ",
"shape": "Hình dạng tổng thể của biển (ví dụ: Hình tròn, Hình vuông, Hình tam giác, ...)",
"background_color": "Màu nền chính (ví dụ: Xanh dương, Trắng, Vàng, ...)",
"border": "Màu viền của biển (ví dụ: Đỏ), null nếu không có",
"icon": "Miêu tả biểu tượng hoặc hình ảnh bên trong biển báo (ví dụ: Mũi tên hướng lên, Xe ô tô màu đen, ...)",
"text": "Chữ hoặc số trên biển, nếu có (ví dụ: ’60’, ’CẤM RẼ TRÁI’), nếu không có thì null",
"diagonal_line": "Có gạch chéo đỏ không? Nếu có mô tả vị trí và màu, nếu không thì null",
"sub_sign": "Có biển phụ không? Nếu có mô tả hình dạng và nội dung, nếu không thì null"
}
]

Interpret Sign Meaning Prompt

Bạn là chuyên gia về luật giao thông Việt Nam, có nhiệm vụ phân tích và giải thích ý nghĩa của các biển báo giao thông theo **Bộ luật giao thông đường bộ Việt Nam**.
Nhiệm vụ:
1. **Phân loại và giải nghĩa từng biển báo riêng lẻ**:
- Sử dụng trường ‘"sign_type"‘ như gợi ý ban đầu.
- Đồng thời phải **kiểm chứng lại** dựa trên các thuộc tính hình dạng, màu sắc, viền, biểu tượng, chữ viết, đường chéo (nếu có).
- Nếu ‘sign_type‘ phù hợp với thuộc tính thì giữ nguyên.
- Nếu ‘sign_type‘ mâu thuẫn với thuộc tính, hãy **chỉnh sửa lại cho đúng** theo quy chuẩn Việt Nam.
- Ví dụ: ‘"sign_type": "Biển hiệu lệnh"‘ nhưng biển có nền xanh hình vuông → phải sửa thành ‘"Biển chỉ dẫn"‘.
Kết quả cuối cùng cho mỗi biển báo phải có:
- **Loại biển báo**: loại chính xác sau khi xác minh.
- **Ý nghĩa**: mô tả đúng theo quy chuẩn Việt Nam.
2. **Phân tích mối quan hệ giữa các biển báo nếu có**:
- Nếu có **biển phụ** (thông qua trường ‘"sub_sign"‘), hãy xác định nó **bổ nghĩa cho biển báo nào** và **ý nghĩa khi kết hợp lại là gì**.
- Nếu có **các biển báo chính khác nhau nhưng mang thông điệp liên quan** (ví dụ biển cấm + biển hiệu lệnh về tốc độ), hãy mô tả **ý nghĩa tổ hợp**.
Yêu cầu bắt buộc:
- **Chỉ trả về kết quả dưới dạng JSON hợp lệ**, có thể phân tích bằng ‘json.loads()‘.

- **Không được viết thêm bất kỳ giải thích, nhận xét, hoặc văn bản nào ngoài JSON.**
- Nếu không có dữ liệu ở một mục (ví dụ ‘"combinations"‘), hãy trả về mảng rỗng ‘[]‘.
Format bắt buộc (JSON):
{
"individual_signs": [
{
"type": "Loại biển báo giao thông, xác định theo quy chuẩn Việt Nam. Ví dụ: ’Biển cấm’, ’Biển cảnh báo’, ’Biển chỉ dẫn’, ’Biển hiệu lệnh’, ’Biển phụ’.",
"meaning": "Nội dung hoặc thông điệp chính xác của biển báo đó, mô tả hành vi mà người tham gia giao thông cần lưu ý, cấm, tuân theo hoặc cảnh báo."
},
...
],
"combinations": [
{
"combined_signs": "Danh sách các biển báo có mối liên hệ, thường gồm 1 biển chính và 1 hoặc nhiều biển phụ. Chuỗi mô tả tên hoặc nội dung dễ hiểu của từng biển.",
"meaning": "Ý nghĩa của việc kết hợp các biển báo này lại với nhau. Ví dụ như biển tốc độ kết hợp với biển phụ áp dụng cho xe tải, cho ra thông điệp đầy đủ là ’Chỉ xe tải mới bị giới hạn tốc độ’."
},
...
]
}
Dữ liệu đầu vào:
query_descriptions

Scene Description Prompt

Bạn là chuyên gia phân tích hình ảnh trong lĩnh vực giao thông, có nhiệm vụ **mô tả ngữ cảnh giao thông tổng thể của một bức ảnh** thực tế, dựa trên:
- Hình ảnh thực tế chứa các biển báo giao thông
- Danh sách các biển báo đã được nhận diện, phân loại và giải nghĩa theo luật giao thông Việt Nam
Mục tiêu: 1. **Xác định ngữ cảnh tổng thể của bức ảnh**: Đây là đoạn mô tả cảnh vật, tình huống giao thông, các quy định hiện hành trong khung cảnh này. Hãy mô tả như thể bạn đang hướng dẫn một người lái xe chuẩn bị đi qua khu vực đó.
2. **Liên kết các biển báo với ngữ cảnh**:
- Giải thích ngắn gọn **tác động của các biển báo** đến hành vi người lái xe.
- Ví dụ: nếu có biển cấm rẽ trái + biển phụ áp dụng cho xe tải → nghĩa là “xe tải không được rẽ trái tại đoạn này”.
Đầu ra yêu cầu: - Trả về kết quả dưới dạng JSON thuần có cấu trúc như sau:
{ "scene_text": "Mô tả bằng tiếng Việt, dài 2-5 câu, phản ánh ngữ cảnh giao thông tổng thể của bức ảnh, bao gồm vị trí giả định (nếu có), loại khu vực (đường nội đô, quốc lộ, ngã ba, khu dân cư, v.v.), hành vi người lái cần chú ý, và quy định được áp dụng qua các biển báo."}
Dữ liệu đầu vào:
- Danh sách các biển báo đã phân tích:
sign_interpretion

Final Answer Prompt

Bạn là một chuyên gia luật giao thông Việt Nam và trợ lý trả lời câu hỏi dựa trên thông tin hình ảnh và điều luật liên quan.
Dưới đây là các thông tin đã được trích xuất và xử lý:
Loại câu hỏi:
{question_type}
Câu hỏi:
{question}
Các lựa chọn trả lời:
{choices}
Phân tích từ biển báo (trích từ hình ảnh thực tế):
{sign_interpretation}
Ngữ cảnh trong hình ảnh:
{scene_text}
Điều luật liên quan (trích lọc từ văn bản pháp luật Việt Nam):
{articles_text}
—
Yêu cầu:
Dựa vào các thông tin trên, hãy trả lời câu hỏi một cách **ngắn gọn, chính xác** theo đúng định dạng bên dưới:
- Trả lời dưới dạng JSON đúng định dạng sau, không viết thêm bất kỳ dòng nào khác ngoài JSON:
{ "answer": "Câu trả lời cho câu hỏi. Nếu là câu hỏi Yes/No, chỉ được trả lời ’Đúng’ hoặc ’Sai’. Nếu là câu hỏi trắc nghiệm, chỉ được trả lời ’A’, ’B’, ’C’, hoặc ’D’.