Prompting Beyond Pixels: A Training-Free and Retrieval-Enhanced Paradigm for Traffic Question Answering

Vu Thanh Dat Ha[1] **Tuan Huy Tran[1]** **Ngoc Chien Chu[1]** **Duc Hao Pham[1]** **Huan Vu[2]**
[1] VNPT AI [2] National Economics University
{thanhdathv, huyttuan, chungocchien, haopd}@vnpt.vn
{huanv}@neu.edu.vn

Abstract

Multimodal legal question answering poses unique challenges: domain-specific terminology, tightly coupled textual and visual evidence, and often limited labeled data. This paper describes our submission to the VLSP 2025 MLQA-TSR competition, which addresses two complementary subtasks: (i) multimodal article retrieval and (ii) multiple-choice question answering grounded in regulatory passages and images. We develop a zero-shot, prompt-driven pipeline (ZIQA) that avoids fine-tuning due to the small training budget. Key components include a preprocessing stage that concatenates and captions images and converts HTML tables to Markdown, a hybrid retrieval system combining visual embedding search and targeted classification, and carefully engineered prompting with heuristic rules to reduce confusion from auxiliary article text. On the public QA split, our selected backbone: **InternVL3-78B with ZIQA+** attains **83.56%** accuracy on private test.

1 Introduction

Multimodal information retrieval and question answering have become central problems in modern vision–language research. In *image retrieval* tasks, the objective is to identify visual items from a large collection that are semantically relevant to a given query (which may be text, an image, or both). Robust retrieval must therefore reconcile differences in representation between modalities while remaining scalable to large repositories. In *image question answering* (Image QA), the goal is to produce an accurate natural-language or discrete answer given an input question and associated visual evidence; success requires both fine-grained visual understanding and the ability to integrate visual cues with textual context.

Both problems pose recurring technical challenges that are amplified in domain-specific settings. First, alignment between visual evidence and textual provisions is non-trivial when regulatory diagrams, captions, and tabular specifications are present. Second, long-form documents and verbose HTML tables can easily exceed model context windows, necessitating careful preprocessing. Third, labeled data are often scarce in specialized domains, which limits the practicality of supervised fine-tuning and motivates zero-shot or few-shot approaches. Finally, multimodal pipelines must be engineered for robustness: noisy auxiliary information (e.g., loosely related “relevant” articles) can act as distractors and degrade performance if not controlled.

This paper addresses these challenges in the context of the VLSP 2025 MLQA-TSR competition. The competition task comprises two complementary subtasks: (i) *multimodal retrieval*, where the system must retrieve the relevant regulatory passages given a question and associated imagery, and (ii) *multiple-choice Image QA*, where a discrete answer is required given the retrieved passages and visual evidence. Each sample includes a mandatory textual passage and may include images and/or tables; samples can reference multiple relevant legal articles, requiring cross-document and cross-modal reasoning.

Faced with limited labeled data and heterogeneous inputs, a zero-shot, prompt-driven design was chosen over fine-tuning. The proposed pipeline combines three key components: (1) *preprocessing* that concatenates and captions images and converts HTML tables to compact Markdown to respect VLM context limitations; (2) a *hybrid retrieval* system that couples embedding-based nearest-neighbor search with targeted classification for domain-critical sign categories; and (3) a *Zero-shot Image QA (ZIQA)* module that performs structured prompt construction (context, question, explicit output rules) and applies a small set of heuristic rules to reduce distractors from auxiliary article text. Production-grade inference frameworks (vLLM, lmdeploy) and high-capacity VLMs (Qwen2.5-VL, InternVL3) were employed to ensure efficient and reliable evaluation.

The main contributions are summarized as follows:

- A practical zero-shot pipeline (ZIQA) for regulatory multimodal QA that emphasizes preprocessing and prompt engineering rather than supervised fine-tuning.
- A hybrid retrieval strategy that balances broad visual/textual matching with deterministic handling of critical sign categories.
- Empirical results on the VLSP competition splits showing strong public-set performance (InternVL3-78B + ZIQA: 83.56% accuracy) and an analysis of failure modes on the more challenging private split.

2 Related Work

2.1 Vision-Language Models

Vision-Language Models (VLMs) have emerged as a central paradigm in multimodal AI, designed to jointly process and reason over visual and textual signals. A seminal contribution is CLIP (Radford et al., 2021), which aligned image and text embeddings through large-scale contrastive pretraining, enabling strong zero-shot generalization. Subsequent systems extended this foundation by incorporating novel architectural modules and training strategies. For instance, BLIP-2 (Li et al., 2023) introduced a Q-former to bridge pretrained vision encoders with large language models (LLMs), MiniGPT-4 (Zhu et al., 2023) leveraged lightweight adapters for efficient alignment, and LLaVA (Liu et al., 2023) exploited gated cross-attention and multi-stage finetuning for robust instruction-following. These models highlight a trend of modular integration, where pretrained vision backbones are coupled with powerful LLMs to achieve advanced visual reasoning and multimodal understanding.

2.2 Multimodal Retrieval

Retrieval plays a crucial role in bridging modalities for practical applications. Early cross-modal retrieval methods relied on handcrafted features (e.g., SIFT for images or keyword search for text), which limited scalability. The advent of contrastive representation learning, particularly with CLIP (Radford et al., 2021), transformed retrieval by enabling direct alignment of heterogeneous inputs in a shared embedding space. BLIP (Li et al., 2023) further enriched retrieval through cross-attention mechanisms that capture fine-grained interactions. More recently, retrieval has been integrated into generative frameworks, giving rise to multimodal retrieval-augmented generation (MRAG). Such approaches extend unimodal RAG (Patrick Lewis, 2020) by grounding LLMs in multimodal evidence to reduce hallucinations and improve factuality. Representative systems include MuRAG (Wenhu Chen, 2022) for efficient nearest-neighbor search and RA-CM3 (Michihiro Yasunaga, 2023) for retrieval-augmented multimodal modeling. This evolution reflects a shift from modality conversion (e.g., pseudo-MRAG converting non-text inputs to text) toward native multimodal pipelines, where MLLMs like GPT-4 preserve information in its original form, enabling richer retrieval and reasoning.

3 Methodology

3.1 Data Description

Let $\mathcal{D} = \{d_j\}_{j=1}^{M}$ be the legal reference database comprising M records. Each record $d_j = (\texttt{law\_id}_j, \texttt{section\_id}_j, \texttt{text}_j)$ is a tuple with the fields defined below.

law_id Canonical short identifier for the legislative instrument (e.g., "36/2024/QH15", "QCVN 41:2024/BGTVT")

section_id Identifier for the unit within the law (e.g., chapter/article/paragraph id).

text The textual content of a legal unit may include inline tags of the form «IMAGE img_id \IMAGE» and «TABLE html_table \TABLE».

3.2 Task Definitions

The VLSP 2025 MLQA-TSR is composed of two complementary subtasks.

Subtask 1 - Multimodal Retrieval: Let (q, I) be an input of the problem, where q is a question of natural language and I is the associated street image. The system shall output a set of reference passages $\hat{\mathcal{R}} \subseteq \mathcal{D}$ relevant to (q, I).

Subtask 2 - Question Answering: Let $(q, I, \mathcal{R})$, $\mathcal{R} \subseteq \mathcal{D}$ is the set of reference passages (terms or article excerpts) provided from the regulatory corpus.

A discrete answer $y \in \mathcal{Y}$ is required. For this task $\mathcal{Y}$ is either the multiple-choice set $\{A, B, C, D\}$ or the binary set $\{\text{Yes}, \text{No}\}$ depending on the question type. Example is shown in Figure 1.

Question: Các loại xe nào được phép lưu thông vào đoạn đường trên trong khoảng từ 6:00 đến 22:00?

A. Xe khách 40 chỗ.

B. Xe ô tô con.

C. Xe đầu kéo.

D. Ô tô kéo rơ moóc.

Reference: Điều 26.1, P.106(a,b) trong Thông tư 54/2019/TT-BGTVT
Correct answer: B

Figure 1: Vietnamese QA instance

3.3 Zero-shot Multimodal Retrieval

In this section, we design a zero-shot multimodal retrieval pipeline that connects traffic sign images with legal documents. As illustrated in Figure 3.1, the process begins with detecting and filtering traffic signs using LLMDet (Fu et al., 2025), followed by classification with Qwen2.5-VL (Bai et al., 2025) to ensure that only valid traffic signs are retained. Relevant cropped images are then matched with queries through a hybrid image retrieval stage, which combines feature-based similarity (via ViTamin embeddings and cosine similarity in Milvus) (Chen et al., 2024) with targeted classification for critical sign categories. Finally, candidate images are linked to legal articles using Qwen3-Embed (Zhang et al., 2025), which aligns references across two subsets of the law database (L1 and L2). This end-to-end approach ensures robust retrieval and precise alignment between visual evidence and textual legal knowledge.

3.3.1 Detecting and Filtering Traffic Signs

In this section, the input images are processed using both a detection model and a vision-language model (VLM).

First, the images from the database and test set are passed through the LLMDet models (Fu et al., 2025) to extract bounding boxes of traffic signs. At this stage, the prompt ‘sign’ is applied to detect and extract all signs present in the image. However, this approach has the drawback that non-traffic signs, such as advertising signs, may also be detected. To address this, all cropped candidate images are subsequently passed through the Qwen2.5-VL-32B-Instruct (Bai et al., 2025) model, which determines whether each image contains a traffic sign. Specifically, the model is guided by visual and color descriptions of general signs and produces a binary answer (‘Yes’ or ‘No’) for each case.

3.3.2 Image Selection

In this stage, given the set of cropped images obtained from the previous filtering step, each question in the test set is paired with its corresponding cropped images and passed through the Qwen2.5-VL-32B-Instruct (Bai et al., 2025) model. The model is tasked with determining whether each image is relevant to the question, producing a binary decision (“Yes” or “No”). In addition, the model is prompted to provide a textual explanation of its choice. Preliminary experiments indicate that the requirement of such explanations improves the accuracy of the model’s decision making, as the reasoning process encourages a more consistent alignment between the question and visual evidence. Importantly, this step is crucial because our objective is to identify the subset of cropped images that are truly pertinent to the question. By retaining only the relevant images, we establish a more precise input for the subsequent image retrieval stage, where these filtered images serve as queries to search for related content within the database.

3.3.3 Image Retrieval

Our method integrates feature-based similarity search with targeted semantic classification to achieve both robust retrieval and precise handling of specific traffic sign categories.

In the first stage, the vitamin_large2_224 (Chen et al., 2024) model is applied to extract feature embeddings from all images in the database as well as from query images. Cosine similarity between embeddings is then used to measure visual closeness, enabling initial retrieval of candidate matches.

In the second stage, we employ Qwen2.5-32B-

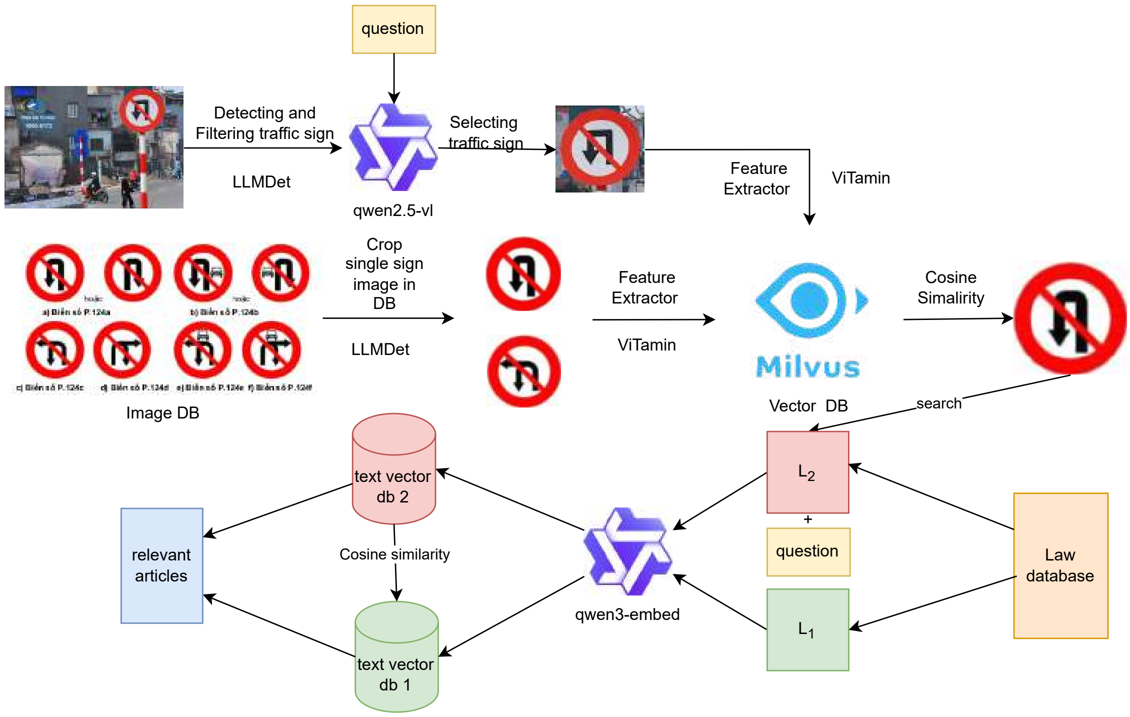


Figure 3.1: Overview of the zero-shot multimodal retrieval pipeline. The workflow integrates traffic sign detection, visual filtering, hybrid image retrieval with cosine similarity, and semantic text retrieval to connect traffic sign images to their corresponding legal articles.

VL (Bai et al., 2025) as a classification module. Each query image is classified into one of three predefined categories: "No Entry sign," "Time sign," or "Other." If the query is classified as No Entry sign or Time sign, we bypass similarity ranking and directly map the query to its corresponding image in the database, ensuring deterministic and interpretable matching for these critical categories. If the query is classified as Other, we retain the cosine similarity results from the first stage to provide the best-matching candidate from the database.

This hybrid approach combines the efficiency of embedding-based retrieval with the reliability of targeted classification, ensuring both broad generalization and accurate handling of domain-specific sign categories.

3.3.4 Text Retrieval

For the database images obtained during the Image Retrieval step, we first searched for articles in the database that explicitly mention the corresponding image identifiers (e.g. image101, image067, …). The database was divided into two subsets: L_1, consisting of legal articles with numeric identifiers (e.g., 1, 2, …), and L_2, consisting of articles with alphanumeric identifiers (e.g., B.2, G.3, …). Since image references were only found in L_2, we restricted the search to this subset. For each candidate passage in L_2, we then retrieved the most semantically relevant article from L_1 by computing cosine similarity over embedding vectors generated by the Qwen3-Embed-8B (Zhang et al., 2025) model. Finally, we applied a post-processing step to eliminate duplicate articles, which yielded the final set of results for Subtask 1.

3.4 Zero-shot Image QA

The Zero-shot Image QA (ZIQA) module is designed to handle passages that contain both textual and visual information, with the goal of producing accurate responses to multiple-choice questions. Since Vision-Language Models (VLMs) are constrained by limited context length, careful preprocessing of multimodal inputs is required. In addition, due to the scarcity of training data, the system adopts a prompting-based approach rather than fine-tuning. Beyond the original question and answer candidates, the organizers also provide retrieved textual and visual information, which serve as supplementary evidence. These retrieved resources, when properly integrated, enhance the model's reasoning capability by compensating for missing context in the original input. An overview of the entire pipeline is shown in Figure 3.2.

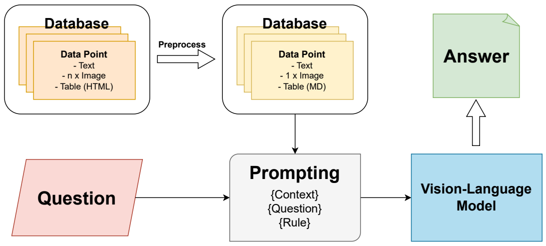


Figure 3.2: Overview of the proposed ZIQA architecture. Raw data points containing text, images, and HTML tables are preprocessed into a unified representation (text, concatenated images, and Markdown tables). The processed passage, together with the input question and handcrafted rules, are composed into prompts for the Vision-Language Model, which generates the final answer.

3.4.1 Data Preprocessing

To ensure compatibility with the context length limitations of the VLM, a series of normalization and compression strategies are applied. An overview of processes is shown in Appendix A.1.

Image Processing Passages may contain multiple images, which are concatenated into a single composite image using the PIL library. Before concatenation, each image is annotated with a caption specifying its identifier in the source passage. This design enables explicit alignment between textual references and visual inputs, while simultaneously reducing the number of images passed to the model. In practice, this mitigates the risk of exceeding the context budget, as a single image typically accounts for 2k–3k tokens.

Table Processing Tables represented in HTML are often verbose and can easily inflate the sequence length. To address this, tables are converted from HTML to Markdown. This conversion reduces the token footprint while retaining the essential structural information, ensuring that the model can interpret the table in a manner comparable to HTML.

3.4.2 Prompting

Given the limited availability of training data, fine-tuning is not performed, as it would likely lead to overfitting and degraded performance. We also considered data augmentation via synthetic data generation. However, this approach requires substantial computational resources for both generation and verification. Human verification would incur significant labeling cost, while automatic verification by models remains unreliable.

Consequently, no model training was performed. Instead, we directly employed pretrained vision-language models and relied on carefully designed prompting strategies to guide the models in solving the tasks. The prompt is structured to include three key elements: (i) the passage context, (ii) the multiple-choice question, and (iii) explicit output formatting rules. This design allows the model to directly generate answers in the desired format, which can then be mapped to the correct choice in the multiple-choice setting, without requiring any additional training.

3.4.3 Model Selection

We adopt Qwen2.5-VL (Bai et al., 2025) and InternVL3 (Zhu et al., 2025) as backbone models due to their leading performance across multimodal benchmarks.

Qwen2.5-VL-72B is the flagship model in the Qwen-VL series. It achieves top-tier results on standard multimodal benchmarks such as MVBench, PerceptionTest, Video-MME, and LVBench—e.g., scoring 70.4 on MVBench and 73.2 on PerceptionTest—surpassing its predecessor Qwen2-VL (Wang et al., 2024), Video-LLaVA (Lin et al., 2023) and GPT-4o (Hurst et al., 2024).

InternVL3-78B is another leading open-source multimodal model, scoring 72.2 on the MMMU benchmark (Yue et al., 2024). It also excels in long-context and domain-specific tasks, demonstrating competitive performance with models like GPT-4o and Gemini-2.5 Pro.

Given the superior performance of larger models on diverse multimodal tasks, we prioritize high-capacity architectures in our pipeline.

4 Experiments

4.1 Dataset

The dataset is constructed from two authoritative legal sources: (i) the National Technical Regulation on Road Signs (QCVN 41:2024/BGTVT), and (ii) the Law on Road Traffic Order and Safety (No. 36/2024/QH15). These documents contain multimodal content, including textual passages, regulatory tables, and illustrative traffic signs.

The dataset is divided into 530 training samples, 50 public test samples, and 146 private test samples. Each sample contains multiple relevant legal articles, and may additionally include images (e.g., traffic signs) and/or tables (e.g., regulatory specifications). The task is framed as multiple-choice question answering, where the model must align textual and visual evidence with the appropriate legal provisions.

4.2 Evaluation Metrics

Subtask 1: Multimodal Retrieval. This subtask is evaluated using the *F2 score*, which places higher weight on recall. For each sample, precision and recall are defined as:

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

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

The F2 score is then computed as:

$$F_2 = \frac{5 \cdot \text{precision} \cdot \text{recall}}{4 \cdot \text{precision} + \text{recall}}.$$

The final score is obtained by averaging the F2 values across all samples.

Subtask 2: Question Answering. This subtask is evaluated using *accuracy*, defined as the proportion of correctly answered questions:

$$\text{Accuracy} = \frac{\#\text{correct answers}}{\#\text{total questions}}.$$

4.3 Implementation Details

Zero-shot Multimodal Retrieval We employed production-level inference frameworks to ensure efficient multimodal retrieval. Specifically, LLmDet was utilized to detect traffic signs as the first-stage filtering mechanism. For visual-language understanding, Qwen2.5-VL-32B was deployed through the vLLM framework to leverage efficient batching and optimized inference. For the image retrieval component, vitamin_large2_224 was adopted, integrated with the Milvus vector database to enable scalable nearest-neighbor search. Finally, for text retrieval, we employed Qwen3-Embed-8B as the embedding backbone to support robust semantic alignment across modalities.

Zero-shot Image QA We employed production-level inference frameworks, namely vLLM and lmdeploy, in order to leverage efficient batching and kernel-level optimizations. Preliminary experiments with vLLM on InternVL3 did not yield satisfactory performance; hence, lmdeploy was adopted for the subsequent experiments with Qwen2.5-VL and InternVL3 models.

To encourage consistent yet informative responses in the sampling regime, we adopted the following generation configuration: $temperature = 0.3$ and $top\_p = 0.95$. All inference experiments were conducted on a computing cluster equipped with 8×NVIDIA A100 40GB GPUs.

4.4 Results

4.4.1 Multimodal Retrieval

We evaluated the proposed multimodal retrieval pipeline on both the public test set and the private test set. The results are reported in Table 1. On the public test, the system achieved strong performance, demonstrating the effectiveness of integrating LLMDet, Qwen2.5-VL-32B, vitamin_large2_224, and Qwen3-Embed-8B. Notably, on the private test, the pipeline maintained consistent accuracy, highlighting the robustness and generalization capability of the proposed approach.

Table 1: Performance of retrieval pipeline on the public and private test set.

Dataset	**F2**
public test	54
private test	58

4.4.2 Question Answering

Public test We evaluated four large-scale models: Qwen2.5-VL-32/72B, and InternVL3-38/78B. Table 2 reports the results on the public test set. Among the baselines, InternVL3-78B achieved the highest accuracy (82%). With the proposed zero-shot prompting strategy (ZIQA), the performance of InternVL3-78B further improved to 84%.

Table 2: Performance of different backbones on the public test set.

Backbone	**Accuracy (%)**
Qwen2.5-VL-32B	64
InternVL3-38B	72
Qwen2.5-VL-72B	78
InternVL3-78B	**82**
InternVL3-78B (ZIQA)	**84**

Private test We observed that the private test set is substantially more diverse and challenging compared to the public test set, which contains only 50 questions with relatively limited variability. The discrepancy between the two distributions, combined with the restriction on the number of submissions, posed significant difficulties for optimizing models without fine-tuning.

Furthermore, we found that providing the model with `reference_article` content occasionally introduced confusion, as the additional information could distract from the actual reasoning process. To mitigate this, we incorporated a set of rules into the prompt design, which improved the model's ability to focus on the essential context and produce more reliable answers called ZIQA+. The detailed rules and prompt examples are provided in Appendix A.2 and A.3. The final results on the private test set are summarized in Table 3.

Table 3: Private Test Accuracy of Different Systems on the Leaderboard.

Team Name	Accuracy (%)
ZIQA+	**83.56**
ZIQA	81.5
InternVL-78B (w/ Image Processing)	79.45
InternVL-78B (w/ Table Processing)	78.76

4.5 Analysis

Preprocessing The two single-component ablations on the InternVL backbone give 79.45% (w/ Image Processing) and 78.76% (w/ Table Processing). Intuitively, this reflects the task characteristics: compact preprocessing of tables and images produces a denser, less noisy context, which improves the model's ability to interpret visual cues such as shape, color, and temporality and to align them with the relevant regulatory text.

Prompting ZIQA attains 81.50% while ZIQA+ reaches 83.56%, an absolute improvement of 2.06 percentage points. This shows that, under a strict zero-shot regime, small but targeted constraints in the prompt can substantially reduce distractors from retrieved passages and steer the VLM to the legally-relevant interpretation of signs and phrases.

Failure-case analysis The example in Appendix A.4 shows an undercount caused by a distant, small-scale car being missed by the detector and thus omitted from the final reasoning pipeline. This is likely due to scale/resolution sensitivity or dataset bias against small instances. Short-term mitigations include multi-scale or sliding-window inference, modestly lowering detection thresholds, and adding automatic crop-and-zoom thumbnails (provided to the VLM) to force attention to small objects. Besides multi-scale inference and crop-and-zoom strategies, an alternative remedy is fine-tuning the vision backbone on traffic-specific data so the model learns to attend to small objects like distant cars. This targeted adaptation can improve robustness in real-world scenarios where vehicles often appear on varying scales.

5 Conclusion

We presented a training-free, retrieval-enhanced paradigm for traffic image question answering, centered on the Zero-shot Image QA (ZIQA) pipeline. By combining careful multimodal preprocessing, a hybrid retrieval stage that fuses visual embeddings with targeted classification, and prompt-engineered zero-shot reasoning, our system attains strong empirical performance on the VLSP MLQA-TSR benchmark—most notably **InternVL3-78B (ZIQA+)** achieving **83.56%** on the private test split. These results demonstrate that, with appropriate retrieval and prompt design, high-capacity vision–language models can deliver competitive results without dataset-specific fine-tuning.

Beyond raw accuracy, our approach emphasizes practicality: (i) preprocessing (concatenation and Markdown conversion) reduces multimodal noise and context bloat; (ii) deterministic handling of critical sign categories improves interpretability; and (iii) a lightweight, prompt-first philosophy reduces development and labeling cost. Together these design choices offer a reproducible baseline for training-free multimodal QA in regulated, image-rich domains.

We acknowledge that the training-free design entails trade-offs. In particular, domain adaptation and fine-grained multi-step reasoning remain challenging under strict zero-shot constraints. Likewise, processing whole passages can introduce redundant context that dilutes model focus. Looking forward, we expect meaningful gains from hybrid directions such as few-shot prompting or compact fine-tuning, incorporation of reranking/chunking to prioritize the most relevant text spans, and efficiency engineering to make the pipeline more suitable for real-time or large-scale deployment.

In sum, ZIQA provides a pragmatic, extensible blueprint for retrieval-augmented, zero-shot multimodal QA. We hope this work serves both as a strong baseline for competitive benchmarks and as a practical reference for applied systems that must reconcile limited labeled data with demanding, domain-specific multimodal reasoning.

Limitations

While the proposed ZIQA framework demonstrates the feasibility of addressing traffic-related image question answering in a training-free manner, two limitations remain noteworthy. First, the system has yet to fully adapt to domain-specific nuances. The scarcity of annotated data in the traffic domain motivates our zero-shot design; however, future work could benefit from more data-efficient strategies such as few-shot prompting or lightweight fine-tuning to enhance domain alignment. Second, our current approach processes entire retrieved passages as context, which may introduce redundancy and dilute the model's focus. A more refined pipeline, for instance incorporating reranking mechanisms to prioritize only the most relevant chunks, could mitigate this issue and further improve answer accuracy.

References

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*.

Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. 2024. Vitamin: Designing scalable vision models in the vision-language era. *Preprint*, arXiv:2404.02132.

Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. 2025. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 14987–14997. Computer Vision Foundation / IEEE.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Weijia Shi Rich James Jure Leskovec Percy Liang Mike Lewis Luke Zettlemoyer Wen-tau Yih Michihiro Yasunaga, Armen Aghajanyan. 2023. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Aleksandra Piktus Fabio Petroni Vladimir Karpukhin Naman Goyal Heinrich Küttler Mike Lewis Wen-tau Yih Tim Rocktäschel Sebastian Riedel Douwe Kiela Patrick Lewis, Ethan Perez. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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*.

Xi Chen Pat Verga William W. Cohen Wenhu Chen, Hexiang Hu. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *arXiv preprint arXiv:2210.02928*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

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*.

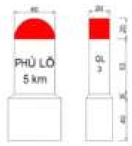
A Appendix

Before Processing	After Processing

Raw context

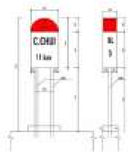
Passages may contain multiple images, each separate.

...- Biển có nền màu xanh lam, viền màu trắng với các đường ôtô khác.



Ghi chú: Kích thước ghi trên hình vẽ là cm

Hình I.1 - Cột kilômét dạng cột thấp



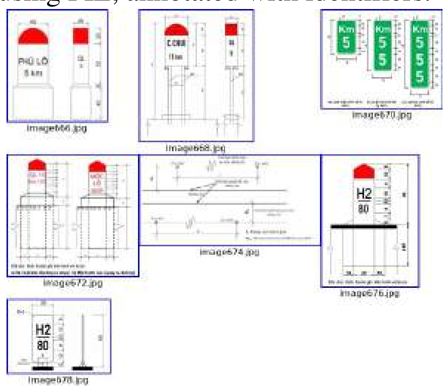
Ghi chú: Kích thước ghi trên hình vẽ là mm

Hình I.2 - Cột kilômét dạng cột cao

...

Processed Context

Images concatenated into a single composite using PIL, annotated with identifiers.



...- Biển có nền màu xanh lam, viền màu trắng với các đường ôtô khác.

<Image Caption: image666.jpg>

Ghi chú: Kích thước ghi trên hình vẽ là cm

Hình I.1 - Cột kilômét dạng cột thấp

<Image Caption: image668.jpg>

Ghi chú: Kích thước ghi trên hình vẽ là mm

Hình I.2 - Cột kilômét dạng cột cao

...

Table Processing (Before)

Tables in verbose HTML format, inflating sequence length.

...

Bảng các giá trị kích thước trên Hình I.3

Đơn vị: mm

«TABLE:

```
<table border=\"0\" cellpadding=\"0\"
      cellspacing=\"0\" class=\"MsoNormalTable\"
       style=\"width:100.0%;border-collapse:
      collapse;mso-padding-alt:0cm_0cm_0cm_0cm\"
       width=\"100%\"> ...
</table>
```

/TABLE»

..

Table Processing (After)

Converted to Markdown, reducing token footprint while retaining structure.

...

Bảng các giá trị kích thước trên Hình I.3

Đơn vị: mm

«TABLE:

| Ký hiệu | A | B | C | D | E | F | G | H | J |

|———–|—|—|—|—|—|—|—|—|—|

| Giá trị kích thước (mm) | 300 | 600 | 12 | 90 | 100 | 70 | 250 | 90 | 40 |

/TABLE»

..

Figure A.1: Overview of Image and Table Processing Pipelines

ZIQA+ prompting strategy

System Prompt: Bạn là một trợ lý ảo có khả năng nhận diện biển báo và hiểu biết luật về đường bộ Việt Nam.

User Prompt:
Dưới đây là một đoạn thông tin trong database:
{concatenated_law}
Suy luận và trả ra đáp án cuối cùng với cụm từ: Đáp án cuối cùng: A, B, C, D... - chỉ được chọn một đáp án.

Một số thông tin cần lưu ý khi trả lời.
{RULE_ADDED}
Nếu câu hỏi không liên quan đến thông tin này thì có thể bỏ qua.

Figure A.2: Prompt templates used in our framework. ZIQA employs the base prompt (black text) without any additional rules. ZIQA+ extends this template by incorporating supplementary rules (highlighted in red) to improve performance under a zero-shot setting.

RULE_ADDED

- Biển báo hình tròn, viền đỏ, nền xanh có một gạch chéo là biển cấm đỗ xe. Cấm đỗ xe thì xe được phép dừng trả khách.
- Biển báo cấm đỗ ngày chẵn, các phương tiện không được đỗ vào ngày chẵn trong tháng (ngày 2, 4, 6,...) tương tự như biển cấm đỗ ngày lẻ(biển có chữ I màu trắng, gạch đỏ chéo) (ngày 1, 3, 5,..).
- Biển cấm có hình: hai ô tô màu đỏ/đen cạnh nhau - hiệu lực cấm tất cả các loại xe cơ giới vượt nhau (kể cả xe được ưu tiên theo quy định) nhưng được phép vượt xe máy 2 bánh, xe gắn máy.

Figure A.3: Supplementary rules incorporated into the ZIQA+ prompting strategy. These rules provide explicit constraints for handling traffic sign interpretation and vehicle regulations, helping the model generate more accurate and context-aware answers.

private_test_135

Input: Dưới đây là một đoạn thông tin trong database

...

a) Để báo đường cấm các loại xe cơ giới kể cả xe máy 3 bánh có thùng đi qua, trừ xe máy 2 bánh, xe gắn máy và các xe được ưu tiên theo quy định, đặt biển số P.103a "Cấm xe ôtô" Nếu đường cấm vì lý do đường, cầu bị yếu, có thể đặt hàng rào chắn hay chướng ngại vật trong phần đường xe chạy.

Hình B.3 - Biển số P.103

...

Trong hình trên có bao nhiêu phương tiện xe ô tô tham gia giao thông vi phạm luật giao thông đường bộ?

A: 1
B: 2
C: 3
D: 4

Output: Trong hình trên, có một biển báo cấm xe ô tô (biển số P.103a). Có một chiếc ô tô đang đi trên đường, vi phạm biển báo cấm này. Các phương tiện khác như xe máy không vi phạm vì biển báo chỉ cấm xe ô tô. Đáp án cuối cùng là A.

Figure A.4: Failure case: the image contains two cars but the model detected only the nearer vehicle; the distant car is much smaller and was missed, resulting in an incorrect prediction (A — 1).