A low-cost low-energy approach to VQA on traffic signs problems

Vu Dinh Anh
FPT IS - AI R&D
Hanoi City, Vietnam
anhvd27@fpt.com

Khiem Vinh Tran
University of Information Technology
Vietnam National University
Ho Chi Minh city, Vietnam
khiemtv@uit.edu.vn

Tran Thi Ha
FPT IS - AI R&D
Hanoi City, Vietnam
hatt64@fpt.com

Abstract

Legal question answering (QA) is gaining attention for its potential to improve access to complex regulations. The VLSP 2025 MLQA-TSR shared task introduces a multimodal challenge that requires interpreting traffic signs through both images and Vietnamese text. It is divided into two subtasks: (1) *image-text retrieval*, which identifies relevant legal references and (2) *visual question answering*, which selects the correct answer from multiple-choice or binary options. We propose a simple retrieval-based pipeline that requires no model training. Text and image features are extracted using Jina Embeddings v3, C-RADIOv2-B, and Owlv2, then stored in Qdrant for cosine similarity search. Retrieved examples directly provide legal terms for subtask 1 and serve as few-shot prompts for Llama 4 Maverick in subtask 2. Our method achieved a top-5 ranking (F2 = 0.54) for retrieval and a top-1 ranking (accuracy = 0.86) for question answering.

1 Introduction

Question answering (QA) is a fundamental problem in artificial intelligence, particularly within natural language processing (NLP), with wide-ranging practical applications. Legal question answering (legal QA) has gained increasing attention for its potential to assist users in retrieving and understanding legal information efficiently. Applications of legal QA systems extend beyond legal aid and include compliance verification, contract analysis, automated legal document review, judicial decision support, regulatory monitoring, public access to legislative information, and emerging areas such as autonomous driving where interpretation of traffic laws and safety regulations is crucial. Road traffic safety regulations are critical for ensuring public safety, and strict adherence to traffic signs is essential for protecting lives and property. Motivated by these concerns, the Shared Task VLSP 2025 MLQA-TSR has been introduced to stimulate NLP research through the QA task, aiming to develop intelligent systems that help users interpret the meanings of road traffic signs and related traffic scenarios, thereby promoting traffic safety awareness.

Question: Xe không được phép chạy bao nhiêu km/h trên đoạn đường này? (What is the maximum speed that vehicles are not allowed to exceed on this road section?)
Subtask 1 - Image-text retrieval
Results: 22, B.27 in QCVN 41:2024/BGTVT
Subtask 2 - Visual question answering
Choices:
A: 70
B: 100
C: 130 ✅
D: 110

Figure 1: A sample training image from VLSP 2025 MLQA-TSR

VLSP 2025 MLQA-TSR consists of two subtasks:

- **Image-text retrieval:** Given an image containing traffic signs and a question written in Vietnamese, the system must identify and return the relevant legal reference(s). These references can be one or more terms [1] from either: the law 36/2024/QH15, or the national technical regulation QCVN 41:2024/BGTVT.
- **Visual question answering:** Given an image with traffic signs, a Vietnamese question,

[1] a term may correspond to an article, clause, or point

and referenced term(s) from the above legal documents, the system must select the correct answer. The answer format depends on the question type and can be either: one choice from four options (A, B, C, D), or a binary judgment (Đúng / Sai).

For example in subtask 1, with Figure 1 and a question "Xe không được phép chạy bao nhiêu km/h trên đoạn đường này?", the system must return article 22 and term B.27 [2] from QCVN 41:2024/BGTVT. In subtask 2, similarly given referenced terms and choices, the system should choose option C. It's easily understanable that these tasks require vast knowledge in both textual and visual aspects. At first glance, object detection model, OCR model, data embedding model(s), a vision-language model are *likely* needed to tackle these tasks.

In this work, we take advantage of both textual and visual representations to effectively perform vector search (Salton et al., 1975). For text, we use Jina Embeddings v3 (Sturua et al., 2024) to generate dense embeddings, while for images, C-RADIOv2-B (Heinrich et al., 2025; Ranzinger et al., 2023) is employed to extract feature vectors. To further enhance visual understanding, Owlv2 (Minderer et al., 2023) is applied to detect traffic signs in images, with each detected object subsequently represented as a vector using C-RADIOv2-B. All embeddings (from training set) are stored and indexed in Qdrant[3] to enable efficient similarity search. During evaluation, each test instance is matched against the database using cosine similarity to retrieve the most relevant examples. The retrieved results are then utilized differently for the two subtasks: for subtask 1 , they provide the relevant legal articles; for subtask 2, they serve as in-context examples for Llama 4 Maverick[4] (few-shots learning (Brown et al., 2020; Alayrac et al., 2022)). Our contributions are following:

- We propose a solution that requires *no training* of models, making it both cost-effective and energy-efficient. The solution architecture is modular, enabling components to be swapped or improved easily (section 4).

- In private testing, our approach achieved top-5 ranking with an F2 score of 0.54 for subtask 1 and top-1 ranking with an accuracy of 0.86 for subtask 2 (section 5, Table 4).

[2] B.27 means appendix B section 27

[3] https://qdrant.tech/

[4] https://www.llama.com/models/llama-4/

2 Related Work

2.1 Legal question answering

Legal question answering (QA) systems (Martinez-Gil, 2023) have attracted significant research attention worldwide due to their potential to provide quick and accurate responses to legal inquiries. These systems typically use a combination of natural language processing (NLP), machine learning, and information retrieval methods to understand legal questions and retrieve relevant legal documents or generate precise answers. Advances in large language models and neural networks have boosted performance, enabling systems to interpret legal contexts more effectively. Global research has focused on handling complex legal language, disambiguating terms, and reasoning over legislative texts and case law (Abdallah et al., 2023). However, challenges remain such as dealing with diverse jurisdictions ensuring explainability of answers, and addressing the evolving nature of laws. Many works emphasize building domain-specific datasets and benchmarks (Fei et al., 2024) to evaluate system accuracy and reliability. Additionally, integrating legal QA systems into real-world applications, such as legal aid and compliance verification, confirms their practical significance worldwide.

In Vietnam, legal QA research is emerging with growing interest in applying AI to support legal advisory services and public administrative processes (Pham Duy and Le Thanh, 2023). Several Vietnamese research groups have built question answering systems tailored to Vietnamese language and legal documents, focusing on improving information retrieval and answer generation under local legal contexts. For instance, recent works (Nguyen et al., 2023a) introduce labeled datasets of legal questions in Vietnamese and propose models that leverage automatic data enrichment and large language models for better performance. Government projects (Vuong et al., 2024) also explore deploying legal QA systems to facilitate citizen access to legal information, enhancing transparency and administrative efficiency. Despite these achievements, challenges remain due to the complexity of Vietnamese legal language and limited availability of digitalized comprehensive legal corpora. Efforts continue to improve system accuracy, expand legal knowledge representation, and ensure language-

specific nuances. Overall, Vietnam is gradually establishing a foundation for effective legal QA systems, contributing to the broader global landscape while addressing unique national needs.

2.2 Visual question answering

Vietnamese Visual Question Answering (VQA) has made significant progress with the introduction of several domain-specific datasets and advances in modeling techniques tailored to Vietnamese language and contexts. The initial milestone in this area was the ViVQA dataset, developed by Tran et al. (2021), which serves as the first large-scale resource specifically designed for Vietnamese VQA. ViVQA (Tran et al., 2021) provides a collection of images paired with relevant natural language questions and answers, thereby establishing a foundational benchmark for evaluating Vietnamese VQA methods. Building upon this foundation, the EVJVQA dataset (Nguyen et al., 2023b) was developed as part of the VLSP 2022 challenge [5]. EVJVQA extends research by providing a multilingual benchmark that includes Vietnamese, English, and Japanese question-answer pairs, encouraging cross-lingual and multilingual approaches in visual question answering systems. Subsequently, OpenViVQA dataset (Nguyen et al., 2023c) was introduced in the VLSP 2023 challenge [6], focusing on more complex, open-ended questions in Vietnamese and enhancing linguistic and contextual diversity.A dditionally, ViOCRVQA dataset (Pham et al., 2025), introduced in 2025, specializes in answering questions about text embedded within images through optical character recognition (OCR), while the ViTextVQA dataset (Van Nguyen et al., 2024), also released in 2025, focuses on large-scale visual question answering involving document images, emphasizing reading comprehension and language understanding in Vietnamese multimedia contexts.

Beyond these, additional datasets such as ViCLEVR (Tran et al., 2024) and LawViVQA (Le et al., 2024) have enriched the research landscape by addressing compositional reasoning tasks and domain-specific applications such as legal visual question answering. LawViVQA, as the first VQA dataset focused on the legal domain in Vietnamese, provides a valuable resource composed of legal document images paired with relevant questions and answers, facilitating research that integrates multimodal understanding with complex legal reasoning. Despite these advances, there remain significant gaps and challenges in the current datasets and modeling approaches, including limited scale and diversity of legal visual data, the complexity of reasoning required for legal interpretation, and the integration of contextual legal knowledge with visual inputs. Motivated by these limitations, a new challenge has been proposed: Multimodal Legal QA on Traffic Sign Rules. This challenge aims to address the complexities of answering legal questions grounded in multimodal data related to traffic regulations, fostering the development of novel AI methods tailored to the intricacies of traffic law interpretation. Our research intends to tackle this challenge by designing models that effectively combine visual and textual legal information to improve accuracy and reliability in the domain.

3 Data Insights

Law corpus is made of 2 law-fundemental documents: the law - 36/2024/QH15 and the national technical regulation - QCVN 41:2024/BGTVT. The law requires the existence of the decrees and the circulars to be particularized. Only those by-law documents instruct both the people and the law enforcement how to deal with certain situations. The national technical regulation acts as the blueprint of roads, signs, and signals (Figure 7). This regulation is surely easily interpreted by technicians and engineers. It could be challenging for the average people to read without prior engineering knowledge. To make law corpus useful both for the humans and the machines, we believe in the presence of decrees, circulars, simplified technical documents, and guideline books by driving schools.

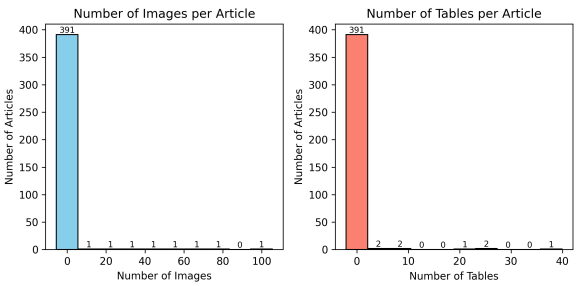


Figure 2: Number of Images and Tables per Article

Most articles in law corpus do not contain images nor tables (Figure 2). Rarely some terms in law corpus have couple tens of images and tables. Most

[5] https://vlsp.org.vn/vlsp2022/eval/evjvqa

[6] https://vlsp.org.vn/vlsp2023/eval/vrc

training samples have less than 3 relevant articles (Figure 3). We found that 40 data points with faulty ‘law_id‘ and ‘article_id‘. For example, ‘article_id‘ is "22.0" but it’s actually "22". These faulty training samples are ignored in any process or pipelines. Images that have width or height bigger than 1536 in pixels are considered high resolution (HR). There are 28 HR images. It’s acknowledged that images have wide range of pixels values (Table 1).

	Min	Max	Average
Width	121	2560	834
Height	126	2560	573

Table 1: Image dimensions in pixels

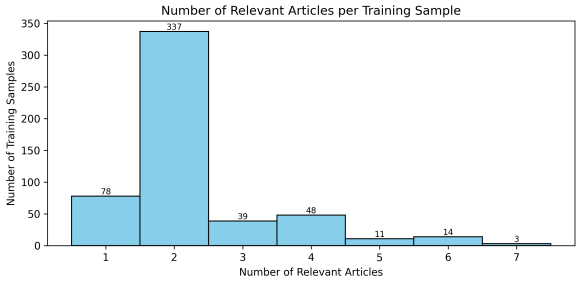


Figure 3: Number of Relevant Articles per Training Sample

4 Methodology

4.1 Vector Indexing

As shown in Figure 4, the system is composed of three parallel data ingestion pipelines. Each data stream is designed to process a specific data modality and embed into a vector representation. It’s the transformation of text (question and choices), image, detected objects into high-dimensional vectors which are indexed in a single Qdrant collection. Notably, only training data is used in these processes.

For the first pipeline, questions and choices are concatenated together. This single string is passed to through Jina Embeddings V3 to produce a vector with dimension of 1024.

The second pipeline is dedicated to image data. The input is the whole image that undergoes a preprocessing step. Images larger than 1536x1536 pixels are resized to these dimensions, either by width or by height. This is because of C-RADIOv2-B can effectively handle images in 32x32 to 1536x1536 pixels. After that, the image is processed by the C-RADIOv2-B model, which transforms it into a 2304-dimensional vector.

In the last pipeline, it also takes an image but performs two steps. The image first fed into Owlv2, an open-vocabulary object detection models. Owlv2 enables the system to detect and identify object within anh image without being limited to a predefined set of categories. Our objects of interests are any traffic signs in any shapes: rectangle sign; triangle sign; square sign; circle sign; octagon sign; stop sign. For each detected object, the C-RADIOv2-B model, the same model used for full image embedding, is utilized to create a new, distinct 2304-dimensional vector(s). Any objects smaller than 32x32 pixels are ignored because it’s out of effective range of C-RADIOv2-B. Each row in the vector database has a list of detect object vectors. This is known as a multi-vector embedding (Khattab and Zaharia, 2020).

In the vector database, text vector and image vector fields are configured to use cosine similarity and HNSW indexing (Malkov and Yashunin, 2020). For detected object vector field, because the list of objects can be none or has very few or has very much, HNSW indexing isn’t enabled to reduce memory usage and speed up vector uploads. MaxSim function [7] (Khattab and Zaharia, 2020) is set as the comparator of a list of object vectors with another the list.

4.2 Image Text Retrieving

Figure 5 shows the vector searching process. The search has 3 steps:

1. getting top n results, by text vector field
2. getting top m results, by image vector field, out of previous step results
3. getting top k results, by detected object vector field, out of previous step results

The text-based filtering is performed first, as it is particularly effective in grounding multiple-choice and yes/no questions, while also accommodating queries expressed with slightly different wordings. Next, entire images are considered, since similar traffic cases typically share similar scenes. Finally, detected traffic signs are compared to identify training samples with most relevant associated articles.

[7] https://qdrant.tech/documentation/advanced-tutorials/using-multivector-representations/

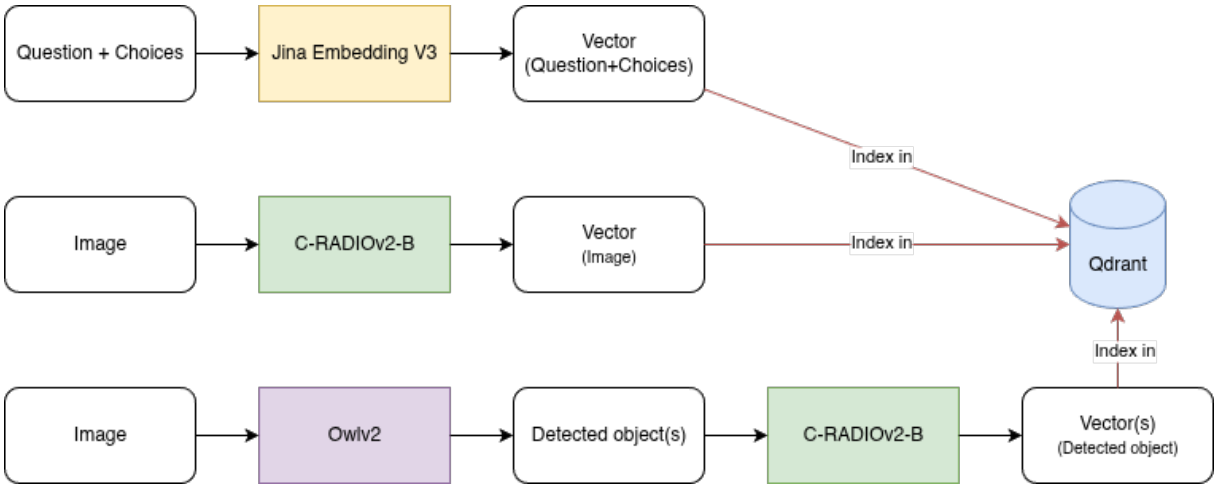


Figure 4: Data feature extractions and data indexing for vector database

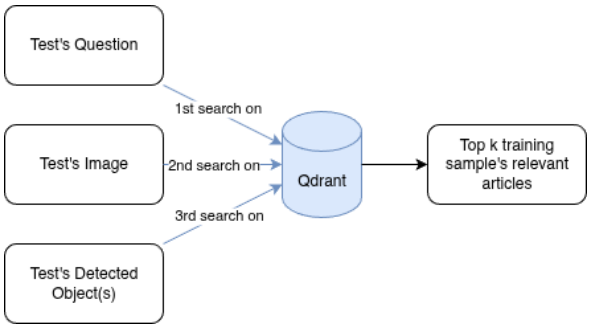


Figure 5: Vector search for subtask 1

Importantly, $k <= m <= n$ must be respected. Through out all experiments, $n = 10$ and $m = 5$ are set while k is in range of 1 to 3. It should be noted that the chosen values of n, m, k were selectively picked. After acquiring closest training samples, we deduplicate relevant articles.

4.3 Visual Question Answering

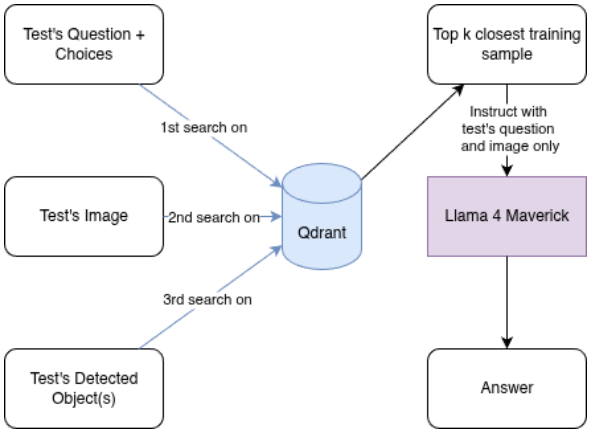


Figure 6: Vector search and few-shots learning with VLM for subtask 2

Llama 4 Maverick is deployed on a server DGX H100 using vLLM (Kwon et al., 2023) with FP8 quantization and max model context length of 131,072 tokens. We acknowledge this is expensive computation because it's a big model in terms of parameters (a 17 billion parameter model with 128 experts) and the hardware is affordable by enterprises. However, with vector search, no training is required, which significantly reduce the cost. The system instruction is written:
« Given an image and a question both about traffic in Vietnam. Multiple choices and yes/no questions shall be provided. If A, B, C, D were given, choose the letter only. If Đúng (Correct); Sai (Wrong) were given, choose Đúng (Correct) or Sai (Wrong) only. No need explanation needed. »

Similarly to previous subsection 4.2, after finding for most-similar training samples, we prepare for few-shots learnings (Figure 6). For each training sample, the prompt is constructed by appending image, question, choices, and the correct choice. After that, the test's image and question are nextly concatenated. The model generates as guided in the system prompt.

5 Evaluation

5.1 Metrics

$$\text{Precision} = \frac{\text{Number of correctly retrieved articles}}{\text{Number of retrieved articles}}$$

$$\text{Recall} = \frac{\text{Number of correctly retrieved articles}}{\text{Number of relevant articles}}$$

$$\text{F2} = \frac{5 \times Precision \times Recall}{4 \times Precision + Recall} \quad (1)$$

$$\text{Accuracy} = \frac{\text{Total correct choices}}{\text{Total questions}} \quad (2)$$

5.2 Results

Top k	F2
1	0.49
2	0.52
3	0.54

Table 2: Private test for subtask 1 using vector search

Top k	Accuracy
1	0.60
2	0.68
3	0.86

Table 3: Private test for subtask 2 using Llama 4 Maverick with few-shots learning

For subtask 1, our vector search pipeline performs 0.54 of F2 on private test. Table 2 shows that the more we retrieve, the higher F2 we get. For subtask 2, with the support of vector search and using 3-shots learning, Llama 4 Maverick can reach 0.86 of accuracy on private test. Table 3 points out a surge in performance while increasing the number of examples for VLM. Combining all results, it proves essentially that a good search system leads to a good QA system.

Comparing the leaderboard results in Table 4 further highlights the strengths and limitations of our approach in both subtasks. For subtask 1, although our vector search pipeline achieves an F2 score of 0.5432, it ranks fifth among participants, trailing behind the top performer with 0.6455. This indicates that while our system is effective, there remains room for improvement in retrieval quality or ranking strategies to boost its competitiveness in this task.

In contrast, subtask 2 demonstrates a significant advantage of our combined vector search and 3-shot learning approach. Our method achieves the highest F2 score of 0.863, outperforming the second-place participant by 0.03. This superior performance underlines the synergy between robust search and advanced learning models in enhancing question-answering accuracy. The consistent lead in subtask 2 confirms that integrating contextual learning with retrieval mechanisms can greatly elevate the system's overall capability.

6 Conclusion and Future Work

In closing, we have reported a *low-cost low-energy* vector search pipeline performing in private test 0.54 of F2 for retrieving information and 0.86 of accuracy for answering multiple-choice questions (VLM with 3-shots learning). This approach has only utilized training dataset, not even law corpus (the law and the technical regulation). Therefore, it's valuable to keep pushing the frontier results (Table 4).

For the future, it is worth exploring *real-time* object detection models, for instance, YOLO-WORLD-V2.1 (Cheng et al., 2024). Only unimodal embedding models are used in this work. Therefore, bimodal image-text embedding models such as SigLIP 2 (Tschannen et al., 2025) should be experimented with for possibly better search results. Future work should also include ablation studies on seperated pipeline components, a detailed analysis of common error cases, and different system instructions for the VLM, which were beyond the scope of this study due to time constraints.

Acknowledgments

The authors would like to thank the following people and entities: people of FPT IS - AI R&D, FPT Smart Cloud.

References

Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. 2023. Exploring the state of the art in legal qa systems. *Journal of Big Data*, 10(1):127.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Participant	F2
Top 1	0.6455
Top 2	0.6114
Top 3	0.5992
Top 4	0.579
Top 5 - Ours	**0.5432**

(a) For subtask 1

Participant	F2
Top 1 - Ours	**0.863**
Top 2	0.8356
Top 3	0.7808
Top 4	0.7329
Top 5	0.726

(b) For subtask 2

Table 4: Leaderboards in private test

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. 2024. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911.

Zhiwei Fei, Xiaoyu Shen, Dawei Zhu, Fengzhe Zhou, Zhuo Han, Alan Huang, Songyang Zhang, Kai Chen, Zhixin Yin, Zongwen Shen, Jidong Ge, and Vincent Ng. 2024. LawBench: Benchmarking legal knowledge of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7933–7962, Miami, Florida, USA. Association for Computational Linguistics.

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA. Association for Computing Machinery.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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. In *2024 RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 393–397. IEEE.

Yu A. Malkov and D. A. Yashunin. 2020. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4):824–836.

Jorge Martinez-Gil. 2023. A survey on legal question–answering systems. *Computer Science Review*, 48:100552.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. 2023. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, pages 72983–73007. Curran Associates, Inc.

Minh Thuan Nguyen, Khanh Tung Tran, Nhu Van Nguyen, and Xuan-Son Vu. 2023a. ViGPTQA - state-of-the-art LLMs for Vietnamese question answering: System overview, core models training, and evaluations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 754–764, Singapore. Association for Computational Linguistics.

Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong TD Vo, Khanh Quoc Tran, and Kiet Van Nguyen. 2023b. Evjvqa challenge: Multilingual visual question answering. *Journal of Computer Science and Cybernetics*, pages 237–259.

Nghia Hieu Nguyen, Duong TD Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2023c. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion*, 100:101868.

Huy Quang Pham, Thang Kien-Bao Nguyen, Quan Van Nguyen, Dan Quang Tran, Nghia Hieu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2025. Viocrvqa: novel benchmark dataset and vision-reader for visual question answering by understanding vietnamese text in images. *Multimedia Systems*, 31(2):106.

Anh Pham Duy and Huong Le Thanh. 2023. A question-answering system for vietnamese public administrative services. In *Proceedings of the 12th International Symposium on Information and Communication Technology*, SOICT '23, page 85–92, New York, NY, USA. Association for Computing Machinery.

Michael Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. 2023. Am-radio: Agglomerative vision foundation model - reduce all domains into one. *Computer Vision and Pattern Recognition*.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv: 2409.10173*.

Khanh Quoc Tran, An Trong Nguyen, An Tran-Hoai Le, and Kiet Van Nguyen. 2021. ViVQA: Vietnamese visual question answering. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 683–691, Shanghai, China. Association for Computational Linguistics.

Khiem Vinh Tran, Hao Phu Phan, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. 2024. Viclevr: A visual reasoning dataset and hybrid multimodal fusion model for visual question answering in vietnamese. *Multimedia Systems*, 30(4):199.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv: 2502.14786*.

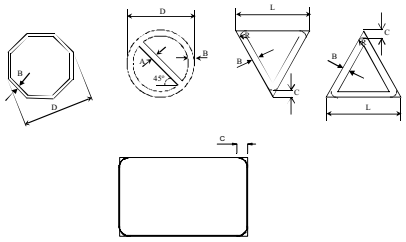
Quan Van Nguyen, Dan Quang Tran, Huy Quang Pham, Thang Kien-Bao Nguyen, Nghia Hieu Nguyen, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2024. Vitextvqa: A large-scale visual question answering dataset for evaluating vietnamese text comprehension in images. *arXiv preprint arXiv:2404.10652*.

Thi-Hai-Yen Vuong, Ha-Thanh Nguyen, Quang-Huy Nguyen, Le-Minh Nguyen, and Xuan-Hieu Phan. 2024. Improving vietnamese legal question–answering system based on automatic data enrichment. In *New Frontiers in Artificial Intelligence*, pages 49–65, Cham. Springer Nature Switzerland.

A Appendix

Figure 7 shows technical drawings and a table detailing the standard dimensions of common traffic sign shapes, including octagonal, circular, triangular, and rectangular forms. Key measurements like outer diameter (D), edge band width (B), side length (L), and corner radius (R) are marked in millimeters to ensure uniformity in sign design.

Additionally, a bilingual Table 5 lists essential legal terms found in law documents, pairing English terms with their Vietnamese equivalents to aid legal professionals and translators. Terms such as "Decree" (Nghị Định) and "Circular" (Thông Tư) refer to formal legal instruments, while organizational units like "Part" (Phần), "Chapter" (Chương), and "Section" (Mục) help structure the documents.



Hình 1 - Kích thước các loại biển báo

Bảng 1 - Kích thước cơ bản của biển báo hệ số 1

Đơn vị tính: mm

Loại biển	Kích thước	Độ lớn
Biển tròn	Đường kính ngoài của biển báo, D	700
	Chiều rộng của mép viền đỏ, B	100
	Chiều rộng của vạch đỏ, A	50
Biển bát giác	Đường kính ngoài biển báo, D	600
	Độ rộng viền trắng xung quanh, B	30
Biển tam giác	Chiều dài cạnh của hình tam giác, L	700
	Chiều rộng của viền mép đỏ, B	50
	Bán kính lượn tròn của viền mép đỏ, R	35
	Khoảng cách đỉnh cung tròn đến đỉnh tam giác cơ bản, C	30
Biển vuông, chữ nhật	Khoảng cách đỉnh cung tròn đến đỉnh chữ nhật cơ bản, C	20-30

13

Figure 7: The 13th page in the national technical regulation - QCVN 41:2024/BGTVT

Code (of Law)	(Bộ) Luật
Decree	Nghị Định
Circular	Thông Tư
Part	Phần
Chapter	Chương
Section	Mục
Article	Điều
Clause	Khoản
Point	Điểm
Appendix	Phụ lục

Table 5: Legal terms related to law documents in English and Vietnamese