A Hybrid Dual-Branch Retrieval and Chain-of-Thought Reasoning Framework for Multimodal Legal Question Answering

Nguyen Van Tien and Mong The Luc and Han Huu Dang
Posts and Telecommunications Institute of Technology
Hanoi, Vietnam
{TienNV.B23KD068, LucMT.B23KH073, DangHH.B23KH019}@stu.ptit.edu.vn

Abstract

This paper presents our comprehensive framework for the VLSP 2025 shared task on multimodal legal question answering, achieving top-tier performance with a 3rd place in Subtask 1 (Retrieval) and a 3rd place in Subtask 2 (Question Answering). Our solution is built on two specialized methodologies. For legal article retrieval, we introduce a novel dual-branch hybrid system that synergizes the strengths of example-based retrieval and dense corpus-based retrieval. This architecture leverages the powerful multimodal embedding capabilities of the **gme-Qwen2-VL-2B-Instruct** model and combines results through a weighted, reranked mechanism, with hyperparameters rigorously optimized for the F2 score. For the question-answering task, we employ a large Vision-Language Model (VLM), Qwen2.5-VL, guided by a meticulously designed Vietnamese Chain-of-Thought (CoT) prompt. This strategy empowers the model to perform robust, step-by-step reasoning by integrating visual information, retrieved legal texts, and the user's query. Our results highlight the effectiveness of combining diverse retrieval strategies with advanced generative models to build a robust legal AI system.

1 Introduction

The domain of Legal Question Answering (LQA) presents a formidable challenge in Natural Language Processing, requiring systems to navigate domain-specific terminology, intricate logical structures, and complex semantics. This challenge is amplified in the context of Vietnamese traffic law, where real-world street images—often noisy and containing irrelevant information—must be interpreted alongside dense legal text. Unlike general-domain QA, LQA demands exceptional factual accuracy, as minor errors can have significant consequences. The VLSP 2025 Challenge on Multimodal Legal Question Answering on Traffic Sign Rules (MLQA-TSR) provides a crucial benchmark for this domain through two interconnected tasks (Nguyen et al., 2025):

- **Subtask 1: Multimodal Retrieval:** Identifying all relevant legal articles from a corpus based on a natural language query and a street-view image.
- **Subtask 2: Legal Question Answering:** Providing a precise answer to a query, given an image and a set of pre-identified relevant legal articles.

In this paper, we detail our system which achieved top-3 ranks in both the retrieval and question answering subtasks. Our primary contributions are threefold:

- We propose a novel dual-branch hybrid retrieval architecture that combines the precision of example-based matching with the broad recall of dense semantic search, a key factor in our high retrieval performance. This design makes our system robust to both common queries similar to the training data and novel, unseen scenarios, a key challenge not fully addressed by single-retriever systems.
- We introduce a Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022) tailored for Vietnamese legal reasoning, which guides a large VLM to produce accurate and interpretable answers without task-specific fine-tuning. This choice reflects a pragmatic strategy for specialized domains where large-scale, annotated datasets are often unavailable, making intelligent prompting a more robust alternative to fine-tuning, which can risk overfitting.
- We implement a robust data preprocessing pipeline, featuring a unique rule-based system

for normalizing and expanding inconsistent legal article identifiers, which proved critical for accurate data mapping.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 details the proposed framework. Section 4 presents our experimental setup and results. Section 5 discusses limitations and future work, and Section 6 concludes the paper.

2 Related Work

2.1 Multimodal Retrieval in Specialized Domains

Multimodal retrieval, the task of finding relevant information from a corpus using a query containing multiple modalities (e.g., text and image), has seen significant advancements with models like CLIP (Radford et al., 2021). A common paradigm is to generate a unified embedding for the multimodal query and perform a dense search. However, some approaches employ a *visual pre-filtering pipeline*, where an object detector or a vision backbone is first used to extract key visual elements before the retrieval step. For instance, one might use an object detector based on DINO-DETR (Zhang et al., 2022) to isolate traffic signs, and then use CLIP to embed only the cropped sign. Our work diverges from this multi-stage pipeline, opting for an end-to-end VLM encoder to create embeddings. We hypothesize this approach is more robust as it avoids potential error propagation from an imperfect object detection stage and allows the model to capture the context of the entire scene, not just the isolated sign.

2.2 Legal Question Answering

Legal Question Answering (LQA) is a challenging subfield of QA that requires high factual accuracy due to the domain-specific terminology, intricate logical structures, and complex semantics of legal texts. Previous works in LQA have explored various techniques, from rule-based systems to deep learning approaches leveraging large language models. (You might add 1-2 specific citations to LQA papers here if you have them, e.g., on Vietnamese LQA or multimodal LQA if available). Our approach specifically tackles the multimodal aspect by integrating visual context into the retrieval and reasoning processes.

2.3 Chain-of-Thought Prompting

Chain-of-Thought (CoT) prompting has proven effective in eliciting complex reasoning from Large Language Models (Wei et al., 2022; Kojima et al., 2022). While generic prompts like "Let's think step by step" are widely used, recent work has shown the benefit of domain-specific CoT structures. Our contribution lies in designing a CoT prompt specifically for the Vietnamese legal context, which mandates explicit reasoning steps such as legal application and evidence description. This structured approach forces the model to ground its answer in the provided legal text, a critical requirement for LQA that is not guaranteed by generic CoT prompts, thereby improving factual accuracy and interpretability.

3 Proposed Framework

Our proposed framework, illustrated in Figure 1, is a two-stage system designed for the retrieval and question answering subtasks. The first stage is a novel dual-branch hybrid retrieval system, which is followed by a Chain-of-Thought based QA system.

3.1 Data Preprocessing

As a foundational step, we applied a meticulous preprocessing pipeline to both the legal corpus and the training data provided by the organizers. This addressed several key challenges in the raw data.

Legal Text Cleaning The raw legal texts contained non-semantic HTML tags, primarily <TABLE:...> and <IMAGE:...>. Unlike approaches that convert tables to natural language, we opted to completely remove them. The rationale is that such tables mostly contain supplementary data not central to the legal reasoning required by the queries.

Article ID Normalization and Expansion This was one of the most critical preprocessing steps. We identified significant inconsistencies in the article_id format, especially within the "QCVN 41:2024/BGTVT" law, with formats like "B.3; 41", "47.15", and "22 B.15". A simple string match would fail to link these questions to the correct articles. To solve this, we developed a rule-based expansion function that:

- Uses a predefined dictionary to map known complex cases (e.g., "47.15" → ["47", "15"]).

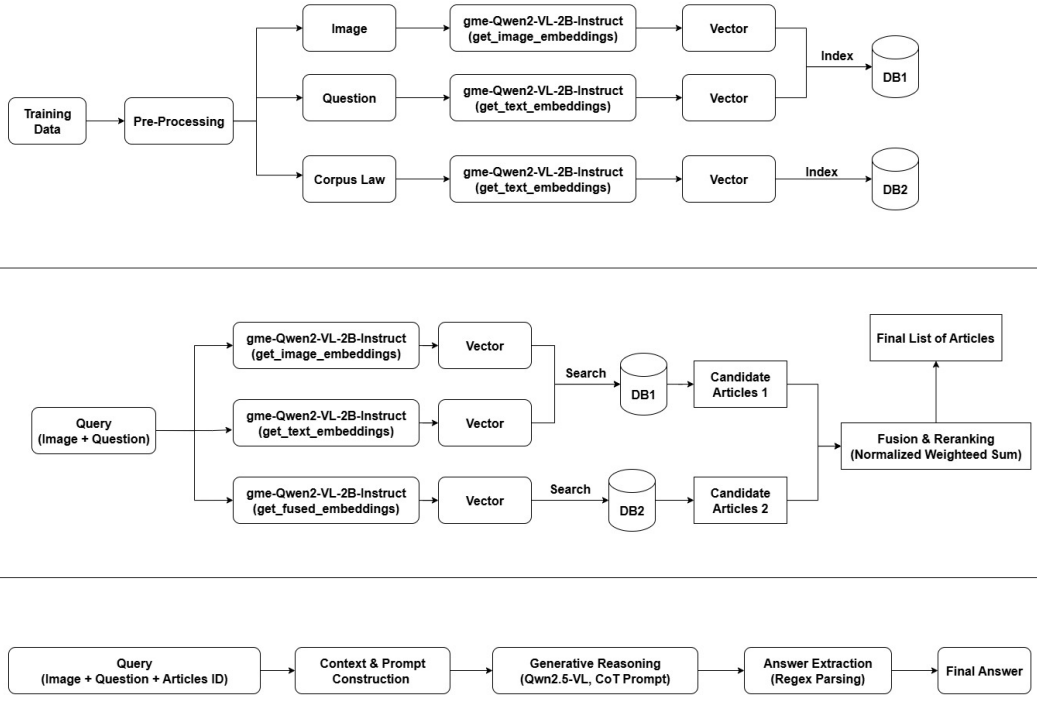

Figure 1: Overall architecture of our proposed framework. (Top) The preprocessing pipeline constructs two indexed databases: DB1 for separate image/text embeddings of training queries, and DB2 for corpus law embeddings. (Middle) The dual-branch hybrid retrieval system encodes a new query into image, text, and fused embeddings, searches DB1 and DB2, and applies weighted fusion with reranking to produce the final list of relevant legal articles. (Bottom) The legal QA stage augments the query with retrieved articles, constructs a Vietnamese Chain-of-Thought prompt, performs step-by-step generative reasoning with Qwen2.5-VL, and applies regex-based answer extraction to yield the final prediction.

- Applies heuristics to split identifiers based on common delimiters such as semicolons, commas, slashes, and spaces.

This normalization process ensured a near-perfect mapping between questions and the legal database, a crucial prerequisite for building our high-performing retrieval system.

3.2 Subtask 1: Dual-Branch Hybrid Retrieval

Our retrieval system processes a multimodal query (image and text) through two parallel branches to leverage different retrieval strategies. We selected **gme-Qwen2-VL-2B-Instruct**[1] as our backbone encoder due to its strong performance in multimodal understanding. While its performance was comparable to CLIP-based models in our initial tests, we opted for the Qwen model family to maintain consistency across both subtasks.

[1] https://huggingface.co/Alibaba-NLP/gme-Qwen2-VL-2B-Instruct

3.2.1 Embedding Strategy

A key design choice is our differentiated use of embeddings for each branch. For Branch 1, which performs example-based matching against the training data, we use **separate embeddings for image and text**. This allows for fine-grained control over their respective contributions via weighted parameters, which is crucial for queries that might be textually similar but visually different to a training example, or vice versa. In contrast, for Branch 2, which performs a direct semantic search against the entire legal corpus, we use a single **combined query embedding**. This holistic representation of the query's semantics is more effective for dense retrieval against a large, general-purpose text index.

3.2.2 Branch 1: Example-Based Retrieval (High Precision)

This branch excels at identifying common or recurring traffic scenarios. It operates on the principle

of case-based reasoning: a new query that is semantically close to a previously seen training query is highly likely to require the same legal articles.

- **Knowledge Base:** We build a knowledge base from the training set, storing separate image and text embeddings for each question, generated by **gme-Qwen2-VL-2B-Instruct**.
- **Weighted Similarity Search:** For a new query, we compute cosine similarities against all entries in the knowledge base. The final similarity score is a weighted combination: $Score = \alpha \cdot Sim_{image} + \beta \cdot Sim_{text}$.
- **Article Aggregation:** We retrieve the articles from the top-k matched training examples, ensuring high relevance for familiar query patterns.

3.2.3 Branch 2: Corpus-Based Dense Retrieval (High Recall)

To handle novel or unseen queries, this branch performs a direct semantic search over the entire legal corpus.

- **Corpus Indexing:** We chunk the entire law database into segments of 512 tokens. Each chunk is embedded and indexed using FAISS (IndexFlatIP) for efficient similarity search (Douze et al., 2024).
- **Combined Query Embedding:** We generate a single, unified query vector using the model's $\text{get\_fused\_embeddings}$ function, which captures the combined semantics of the image and question text.
- **FAISS Search:** The combined embedding is used to search the index, retrieving the most semantically relevant text chunks from the entire law database, ensuring broad coverage.

3.2.4 Optimal Fusion and Reranking

The key to our system's success lies in the fusion strategy. The raw scores from both branches are not directly comparable.

- **Score Normalization:** We apply min-max scaling to the scores from each branch independently.
- **Weighted Fusion:** The final score for each candidate article is a weighted sum: $Final\_Score = w_1 \cdot Norm\_Score_{B1} + w_2 \cdot Norm\_Score_{B2}$.
- **Hyperparameter Optimization:** The weights $(\alpha, \beta, w_1, w_2)$ and the final cutoff N are not chosen arbitrarily. We performed an extensive, nested grid search on a held-out validation set to find the optimal parameter combination that maximized the F2 score. This meticulous, data-driven optimization process was instrumental in elevating the system from a simple heuristic combination to an empirically optimized ensemble, a critical step in achieving our top-3 result.

3.3 Subtask 2: Context-Augmented Generative QA System

For Subtask 2, our philosophy was to rely on the advanced reasoning capabilities of a large, powerful VLM rather than a complex, multi-stage filtering pipeline. This approach simplifies the system architecture while leveraging the cutting edge of generative AI.

3.3.1 Model Choice and Context Strategy

We selected Qwen2.5-VL, a larger and more capable model than many compact alternatives (Qwen Team, Alibaba Group, 2025). We hypothesized that its superior reasoning and context comprehension abilities would allow it to effectively process a less filtered, more comprehensive context. Our strategy involves concatenating the full, preprocessed text of all provided relevant articles. This direct approach trusts the model's attention mechanism to identify the most salient information, eliminating the need for intermediate re-ranking or chunk validation steps.

3.3.2 Vietnamese Chain-of-Thought (CoT) Prompting

Our main innovation in this task is the design of a structured, Vietnamese-language Chain-of-Thought prompt. This prompt does not simply ask for an answer; it commands the model to externalize its reasoning process, significantly improving reliability and accuracy, drawing inspiration from the foundational work on CoT prompting (Wei et al., 2022). Our approach is a form of zero-shot reasoning, adapting the principles of prompts like "Let's think step by step" (Kojima et al., 2022) to the specific structure of Vietnamese legal analysis. The creation of a prompt in Vietnamese is a non-trivial contribution, representing not just a translation but a linguistic and cultural adaptation of a reasoning framework, thereby contributing to

the underexplored area of multilingual prompt engineering. The full prompt templates are provided in Appendix A. The prompt template is structured with the following components:

- **Role Instruction:** "You are a Vietnamese traffic law expert."
- **Provided Legal Context:** The concatenated text of the relevant articles.
- **Structured Reasoning Steps:** A mandatory template for the model to complete:
  - Step 1 - Describe the image
  - Step 2 - Apply the law from the context
  - Step 3 - Reasoning to link image and law
  - Step 4 - Conclude and state the final answer (for multiple choice) or Determine True/False (for yes/no).
- **Final Answer Placeholder:** A clear format {{ĐÁP ÁN: X}} or {{ĐÁP ÁN: Đúng/Sai}} to ensure parsable output.

This CoT structure forces the model to deconstruct the problem, analyze the visual evidence, ground its reasoning in the provided legal text, and synthesize a conclusion, mirroring an expert's workflow.

3.3.3 Robust Answer Extraction

The final step is a robust parsing function that uses a series of prioritized regular expressions to extract the answer from the model's verbose CoT output, ensuring compliance with the required submission format.

4 Experiments and Results

4.1 Experimental Setup

Dataset and Validation Split We used the official MLQA-TSR dataset provided by the VLSP 2025 organizers (Nguyen et al., 2025). For hyperparameter tuning and model selection, we created a validation set by randomly splitting the official training data into a 90% training partition (477 samples) and a 10% validation partition (53 samples). For the final submission on the private test set, the knowledge base for our example-based retrieval branch was built using the full training dataset (530 samples).

Method	Acc@1	Acc@5	Acc@10	Acc@20
Text-only Baselines				
BM25	10.0	28.0	34.0	38.0
Qwen3-Embedding-8B	14.0	38.7	49.6	67.9
Multimodal Model				
gme-Qwen2-VL-2B-Instruct	18.87	50.57	64.72	78.30

Table 1: Preliminary retrieval model comparison on the validation set using Accuracy@k (%).

Evaluation Metrics As per the competition rules, the official evaluation metric for Subtask 1 (Retrieval) is the mean F2 Score, which balances precision and recall with a higher weight on recall. For Subtask 2 (Question Answering), the metric is Accuracy (Nguyen et al., 2025).

Hardware All development experiments were conducted on Kaggle notebooks equipped with NVIDIA T4 GPUs (16GB VRAM). Due to the significant memory requirements of the largest model, the final inference for Subtask 2 using the Qwen2.5-VL-72B model was performed on a rented cloud server with high-performance GPUs.

Implementation Details Our system was implemented using Python with the PyTorch framework (Paszke et al., 2019). We utilized the Hugging Face Transformers library for model access (Wolf et al., 2020) and FAISS for efficient similarity search (Douze et al., 2024). Key models included **gme-Qwen2-VL-2B-Instruct** for retrieval and the Qwen2.5-VL series for question answering (Qwen Team, Alibaba Group, 2025).

4.2 Subtask 1: Retrieval Performance

Our retrieval experiments were conducted in two phases: an initial exploration to select the best backbone model, followed by a rigorous ablation study of our proposed hybrid system.

4.2.1 Preliminary Model Exploration

To select the most effective embedding model, we conducted a series of preliminary experiments comparing traditional, text-only, and multimodal approaches. During this rapid development phase, we used Accuracy@k (Acc@k) on our validation set for evaluation. The results are shown in Table 1.

The results in Table 1 reveal a clear performance trend. Dense retrieval models like Qwen3-Embedding significantly outperform the sparse BM25 baseline. More importantly, the multimodal **gme-Qwen2-VL-2B-Instruct** model, which processes both image and text, demonstrates a substantial improvement over the text-only dense retriever

Method	F2 (Train)	F2 (Public)
Alternative Pipeline (DINO + CLIP)	0.16	-
Our System Components		
Branch 1 (Example-Based Only)	0.60	0.40
Branch 2 (Corpus-Based Only)	0.27	0.22
Our Hybrid System (Final Weighted Combination)	**0.70**	**0.56**

Table 2: Ablation study of our retrieval system components using the official F2 Score.

Model	Context Provided?	Method / Note	Acc (%)
Vintern-3B-R-beta	No	Baseline	58
Qwen2-VL-3B	No	Baseline	60
Qwen2.5-VL-7B	Yes	CoT Prompting + Truncate	75

Table 3: Performance of different VLM configurations for Question Answering.

across all top-k metrics. This confirmed our hypothesis that visual information is critical for this task. Based on these findings, we selected **gme-Qwen2-VL-2B-Instruct** as the core model for both branches of our final system.

4.2.2 Ablation Study and Hyperparameter Optimization

We performed a rigorous ablation study to validate our dual-branch hybrid architecture, using the official F2 score. The results are in Table 2.

The study in Table 2 provides several key insights. First, our exploration of a complex pipeline using an object detector (DINO (Caron et al., 2021)) to pre-process images yielded a very low F2 score of 0.16. This result suggested that such a multi-stage approach was prone to error propagation and less robust than an end-to-end multimodal model for this specific task. Second, within our proposed system, Branch 1 (Example-Based) served as a strong component, achieving a 0.40 F2 score on the public test. Branch 2 (Corpus-Based), while having a lower standalone F2 score, was critical for providing high recall on novel queries not well-represented in the training data. The most crucial finding is the significant performance gain from our weighted combination mechanism. The complete hybrid system boosted the F2 score to 0.56 on the public test set, a 40% relative improvement over the strongest single branch's score of 0.40. This result unequivocally demonstrates the powerful synergy between our high-precision, example-based branch and our high-recall, corpus-based branch. Furthermore, we conducted a grid search for the optimal fusion weights $(\alpha, \beta, w_1, w_2)$. The parameters α (for image similarity in Branch 1) and w_1 (for Branch 1's score in final fusion) were searched in the range [0.1, 0.9] with a step of 0.1, while β and w_2 were set to $1 - \alpha$ and $1 - w_1$ respectively. The optimal parameters were found to be $\alpha = 0.6$, $\beta = 0.4$, $w_1 = 0.6$, and $w_2 = 0.4$. This data-driven optimization yielded an F2 score improvement of approximately 0.3 points on our validation set compared to a baseline with uniform weights (all 0.5), demonstrating the effectiveness of our fusion strategy.

4.3 Subtask 2: Question Answering Performance

For the QA task, our experiments focused on assessing the reasoning capabilities of different VLM models and the impact of our Vietnamese Chain-of-Thought (CoT) prompt. Experiments were conducted on our development set (10% of the official training data).

As shown in Table 3, baseline models without legal context achieved accuracies around 58-60%. This indicates a limited ability to answer correctly from the image and question alone. Crucially, when providing the full legal context, the smaller Qwen2-VL-3B model encountered an out-of-memory error on our NVIDIA T4 GPU setup (16GB VRAM), highlighting the necessity of larger models with sufficient context windows or more efficient memory management for this task. We therefore transitioned to the Qwen2.5-VL-7B model, which successfully incorporated the legal context. The combination of this model with our meticulously designed Vietnamese CoT prompt (see Appendix A) and tuning of generation hyperparameters (temperature, top_p) enabled us to achieve a strong accuracy of 75% on our development set. To maximize performance for the final competition phase, we employed the largest available model, Qwen2.5-VL-72B. Due to significant computational and time constraints, this model was used directly for the private test set submission on a rented cloud server. This strategic decision was validated by the final result of 78% accuracy.

4.4 Final Official Results

Our final systems were evaluated on the blind private test sets, achieving top-3 ranks in both subtasks. The official results provide the ultimate validation of our framework.

Task	Final Score	Final Rank
Subtask 1: Retrieval	0.59 (F2 Score)	3rd
Subtask 2: QA	78% (Accuracy)	3rd

Table 4: Final official results on the VLSP 2025 private test set.

5 Limitations and Future Work

Despite its strong performance, our framework has several limitations that offer clear avenues for future research.

Granularity in Visual Information Processing Our final retrieval model processes the entire image as a single input. While effective, this approach lacks granular understanding of specific visual elements. Initial experiments involved a more complex *visual pre-filtering pipeline*. In this setup, we first used a DINO-based model (Caron et al., 2021) to crop traffic signs from the images. The cropped sign image was then passed to a CLIP-based encoder to generate a visual embedding. These embeddings were then combined with text embeddings for retrieval. This approach yielded a very low F2 score (0.16) in our initial trials. We attribute this failure to high variance in image quality, occlusions, and the potential semantic loss from ignoring the surrounding visual context (e.g., other vehicles, road markings) which is crucial for legal reasoning in traffic scenarios. This led us to pivot to our final, more direct dual-branch architecture. Future work should revisit this by training a domain-specific traffic sign detector to isolate pertinent visual evidence more reliably.

Dependency on Training Data Distribution The high performance of our example-based retrieval branch (Branch 1) indicates a significant dependency on the training dataset. While effective for queries semantically similar to the training distribution, its ability to generalize to entirely novel multimodal scenarios is inherently limited. The corpus-based branch mitigates this to an extent, but the system's overall balance is still skewed towards known patterns.

Underutilized Data Sources Our current data preprocessing pipeline discards valuable structured information. Specifically:

- **HTML Tables:** Legal texts containing tables with technical specifications (e.g., sign dimensions) were stripped of this content. Future work could parse these tables into a structured format (e.g., key-value pairs) and append them as metadata to the text chunks, providing the models with richer, more precise context.
- **Canonical Law Images:** The legal documents contain clean, canonical images of each traffic sign. These were not utilized in our pipeline. These images could be used to build a powerful visual knowledge base, enabling more accurate matching between noisy real-world signs and their official definitions.

The Strategic Choice Against Fine-tuning We deliberately opted against fine-tuning our models. Domain-specific fine-tuning for legal AI requires a substantial volume of high-quality annotated data to be effective without risking overfitting or catastrophic forgetting. Given the dataset's scale, we concluded that a zero-shot, prompt-based approach was a more robust and pragmatic strategy. This self-criticism highlights a broader challenge and a key takeaway: for many real-world applications where perfect, large-scale data is unavailable, the "model steering" accomplished via advanced prompting can be a more robust and accessible strategy than the "model reshaping" that occurs during fine-tuning.

Constrained Exploration of Vision-Language Models For Subtask 2, our exploration was largely confined to the Qwen series of models (Qwen Team, Alibaba Group, 2025). To achieve our final accuracy, we found it necessary to scale up to the computationally intensive 72B parameter model. This stands in contrast to other top-performing teams that achieved excellent results with significantly smaller models, such as the 11B parameter LLaMA 3.2-Vision. This suggests that our Chain-of-Thought prompt, while effective for Qwen, may not be universally optimal. Future efforts should involve a broader evaluation of different VLM architectures and a focus on developing more efficient prompting techniques that can elicit strong reasoning from smaller, more deployable models.

6 Conclusion

In this paper, we presented a high-performing, dual-system framework for the VLSP 2025 multimodal legal QA challenge, securing 3rd place in both retrieval and question answering. Our work makes two key contributions. First, we demonstrate

the superiority of a hybrid, dual-branch retrieval system that strategically combines example-based and dense corpus-based methods, with its effectiveness maximized through data-driven hyperparameter optimization. Second, we showcase the power of a Vietnamese-centric Chain-of-Thought prompting strategy to guide a large VLM towards accurate and interpretable legal reasoning without requiring any task-specific fine-tuning. Our results validate the philosophy of leveraging intelligent system design and advanced prompting as a robust and effective alternative to fine-tuning in specialized, low-data domains.

References

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660.

Matthijs Douze, Anton Guzhva, Cheng Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazar'e, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems 35*, pages 22199–22213. Curran Associates, Inc.

Minh-Luan Nguyen, Ngan L.T. Nguyen, Kiet Van Nguyen, Van-Hiep Tran, Tuan Vo, Son T. Luu, Hieu Nguyen, and Ngan Tran. 2025. Vlsp 2025 challenge on multimodal legal qa on traffic sign rules. In *Proceedings of the 11th International Workshop on Vietnamese Language and Speech Processing (VLSP 2025)*. Association for Computational Linguistics. (Forthcoming).

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Qwen Team, Alibaba Group. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35*, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Philipp von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Hao Zhang, Feng Li, Wenbo Liu, Hanzi Li, Guodong Wang, Zhaohui Wang, Jianming Xu, and Jun Liu. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

A Prompt Templates for Subtask 2

We used two distinct Chain-of-Thought prompt templates tailored to the question format.

A.1 Template for Multiple Choice Questions

Bạn là chuyên gia luật giao thông đường bộ Việt Nam. Phân tích câu hỏi theo các bước sau:
NGỮ CẢNH PHÁP LÝ:
{context}

HÌNH ẢNH:
Quan sát kỹ hình ảnh để hiểu đầy đủ tình huống giao thông.

CÂU HỎI
{question_text}

LỰA CHỌN: {choices_text}

PHÂN TÍCH TỪNG BƯỚC:
Bước 1 - Mô tả hình ảnh: [Mô tả ngắn gọn những gì quan sát được]
Bước 2 - Áp dụng luật: [Xác định điều luật/ quy định cụ thể từ ngữ cảnh]
Bước 3 - Suy luận: [Giải thích tại sao lựa chọn này đúng]
Bước 4 - Loại trừ: [Giải thích tại sao các

lựa chọn khác sai]

{{ĐÁP ÁN: X}}

A.2 Template for Yes/No Questions

Bạn là chuyên gia luật giao thông đường bộ Việt Nam. Phân tích câu hỏi theo các bước sau:

NGỮ CẢNH PHÁP LÝ:
{context}

HÌNH ẢNH:
Quan sát kỹ hình ảnh để hiểu đầy đủ tình huống giao thông.

CÂU HỎI:
{question_text}

PHÂN TÍCH TỪNG BƯỚC:
Bước 1 - Mô tả hình ảnh: [Mô tả ngắn gọn những gì quan sát được]
Bước 2 - Áp dụng luật: [Xác định điều luật/ quy định cụ thể từ ngữ cảnh]
Bước 3 - Suy luận: [Giải thích logic dẫn đến kết luận]
Bước 4 - Kết luận: [Xác định đúng hay sai dựa trên luật pháp]

{{ĐÁP ÁN: Đúng/Sai}}