ChartTableViVQA: Vietnamese Visual Question Answering for Chart and Table Images with Instruction

Quang Nhat Truong[1,2], Son Thanh Luu[1,2]

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{22521207}@gm.uit.edu.vn
{sonlt}@uit.edu.vn

Abstract. The rapid advancement of Visual Question Answering (VQA) systems has unlocked significant potential for processing chart and table image data. However, in Vietnam, the scarcity of Vietnamese-language VQA datasets poses a substantial barrier to developing multimodal AI models tailored to local linguistic and cultural contexts. To address this challenge, we introduce ChartTableViVQA, a Vietnamese dataset comprising 395 chart images and 474 table images sourced from prominent websites, accompanied by 13,069 self-generated question-answer pairs. The questions span diverse categories, including calculations, comparisons, and trend analysis, designed to evaluate the models' ability to interpret visual content and semantic information in multimodal VQA tasks. The dataset is benchmarked using the state-of-the-art Qwen2.5-VL model, renowned for its superior visual perception and natural language processing capabilities. Experimental results demonstrate that Qwen2.5-VL achieves high performance across key question categories, underscoring the efficacy of ChartTableViVQA in enhancing Vietnamese VQA systems. This dataset serves as a valuable resource for advancing Vietnamese VQA research and lays foundational groundwork for future studies on Vietnamese VQA models, while offering opportunities for scalability across diverse domains and languages.

Keywords: Visual Question Answering · Vietnamese · Chart Images · Table Images

1 Introduction

The emergence of Artificial Intelligence (AI) has significantly advanced Visual Question Answering (VQA) systems, enabling effective interaction with visual content such as charts and tables. These systems are crucial for interpreting complex visual data and extracting meaningful insights. In Vietnam, where visual data representations like charts and tables are widely used, there is an increasing demand for VQA tools that align with the Vietnamese language and cultural context. However, the development of such tools is hindered by the scarcity of high-quality Vietnamese-language VQA datasets.

A primary challenge is the lack of Vietnamese VQA datasets focused on chart and table images, which are critical for various applications. Most existing VQA research focuses on widely spoken languages like English, resulting in multimodal AI models that are less effective for Vietnamese, a language with unique grammatical and cultural characteristics. Additionally, the creation of diverse questions and paraphrased answers to enhance model generalization remains underexplored in current datasets.

This study aims to address these gaps by introducing ChartTableViVQA, a preliminary Vietnamese VQA dataset focused on chart and table images. The objective is to provide an initial resource for developing multimodal VQA models tailored to the Vietnamese context, incorporating varied questions related to calculations, comparisons, and trend analysis. Furthermore, this work investigates the use of paraphrased answers to improve model performance and generalization. By establishing this foundation, the study seeks to support future advancements in Vietnamese VQA research, offering initial insights and resources to guide subsequent developments.

This study offers three contributions to the field:

1. The creation of ChartTableViVQA, a preliminary Vietnamese VQA dataset for chart and table images, addressing a critical gap in resources. This dataset comprises 395 chart images and 474 table images, sourced from prominent Vietnamese websites, reflecting real-world visual data. It includes 13,069 self-generated question-answer pairs, spanning diverse categories such as calculations, comparisons, and trend analysis, designed to evaluate models' ability to interpret visual content and semantic information in multimodal VQA tasks. The dataset also incorporates paraphrased answers for a subset of questions to support model generalization, serving as a foundational resource for advancing Vietnamese VQA systems.
2. An evaluation of existing VQA models to provide reference points for future Vietnamese VQA research.
3. An analysis of errors to explore how training with paraphrased answers may enhance model generalization compared to single-answer approaches.

The paper is organized as follows: **Section 2** reviews related **work. Section 3** describes the construction of the ChartTableViVQA dataset, including data collection, preprocessing, and paraphrased answer generation. **Section 4** outlines the research methodology and experimental setup. **Section 5** presents the results and error analysis. **Section 6** concludes with implications and suggestions for future research directions.

2 Related Work

2.1 Vision Table Question Answering

Vision Table Question Answering (Vision Table QA) focuses on answering natural language questions based on table data presented in the form of images,
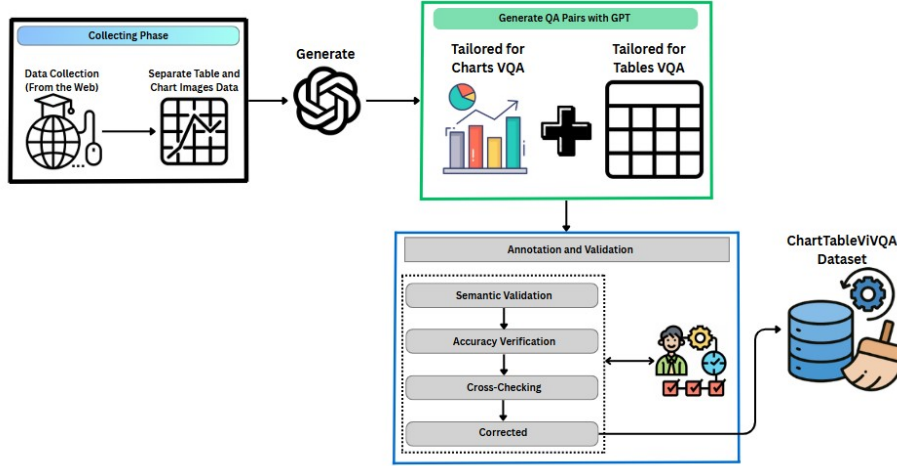
Figure 1: *ChartTableViVQA Dataset Creation Workflow*

such as scanned documents or screenshots. Recently, new datasets have been developed to support Vision Table QA research. For example, TableVQA-Bench [1] provides a dataset with 1,500 question-answer pairs, generated from existing table recognition and table question datasets using large language models to generate questions. Similarly, WikiDT [2] offers a large-scale dataset with 70,919 images and 159,905 tables, supporting research on table recognition and question answering. Additionally, the MMTab dataset [3] is a comprehensive resource for multimodal table understanding, consisting of 97,000 images for pre-training, 82,000 images with instructions and answers, and 23,000 images for model evaluation, covering various table-related tasks. Other important datasets include PubTables-1M [4], with over 1 million tables from scientific documents, Table-Bank [5], a large dataset for table detection and recognition based on images, and FinQA [6], focusing on financial tables. These datasets enable researchers to evaluate model performance on tasks such as table detection, structure recognition, and classification. Major challenges in Vision Table QA include handling diverse table layouts, such as tables with cells spanning multiple rows or columns, and answering complex questions that require logical reasoning or arithmetic operations.

2.2 Vision Chart Question Answering

Vision Chart Question Answering (Vision Chart QA) focuses on answering natural language questions about data visualizations, such as bar charts, line graphs, and pie charts, requiring both visual perception and logical reasoning. An important benchmark in this field is ChartQA [7], which includes 9,600 human-written questions and 23,100 questions generated from human-written chart

summaries, emphasizing visual and logical reasoning on charts. Proposed models for ChartQA include transformer-based architectures that combine visual features from chart images with underlying data tables to answer questions accurately. LEAF-QA [8] introduces a dataset with 250,000 annotated chart images and approximately 2 million question-answer pairs, focusing on real-world scientific charts. ChartQA-X [9] extends ChartQA by providing detailed explanations for the answers, enhancing the model's ability to convey complex information. The dataset includes 28,299 questions, answers, and explanations related to various types of charts, allowing for improved understanding of how answers are derived from the visual data in charts. This approach aims to improve the interpretability and trustworthiness of models performing visual reasoning on charts.

2.3   Models for Vision Question Answering

Qwen is a series of large multimodal models developed by Alibaba Cloud for natural language and visual data processing. The latest model, Qwen2.5-VL, released in January 2025, features up to 72 billion parameters and achieves better performance than models like GPT-4o and Claude-3.5 Sonnet on multimodal tasks, especially in image-based question answering [10]. Qwen2.5-VL combines a Vision Transformer (ViT) with the Qwen2.5 language model, using advanced techniques like dynamic resolution processing and multimodal position encoding (M-RoPE). These improvements enable the model to handle images of varying sizes and scales while maintaining strong performance in visual comprehension and reasoning [10]. The model is trained on a dataset of 4.1 trillion tokens, including image captions, OCR data, visual knowledge, and chart and table data. It is specifically trained on 1 million synthetic chart samples and 6 million real-world table samples, enhancing its ability to analyze charts and extract structured data from tables [10]. Qwen2.5-VL excels in tasks involving document and chart understanding, outperforming other models in benchmarks like OCRBench_v2. It also supports multilingual text recognition, including Vietnamese, in images, making it a versatile tool for complex, multimodal, and multilingual applications [10].

3   Dataset

The **ChartTableViVQA** dataset is meticulously crafted to drive advancements in Visual Question Answering (VQA) for charts and tables in the Vietnamese language. Comprising images of charts and tables sourced from authoritative online platforms, the dataset is enriched with questions and answers automatically generated via the GPT language model. These questions and answers undergo rigorous editing to ensure accuracy and relevance before being validated by a team of annotators to uphold quality standards.

The **ChartTableViVQA** dataset serves as a critical resource for researchers developing VQA models, significantly contributing to progress in natural language processing and computer vision. It provides a robust platform for testing

and refining VQA algorithms, particularly in the Vietnamese context, where high-quality datasets remain scarce. By enabling the construction of models capable of interpreting and responding to complex questions based on visual data, the dataset unlocks potential applications across interdisciplinary domains.

3.1   Data Collection

The construction of the **ChartTableViVQA** dataset commenced with the acquisition of chart and table images from leading Vietnamese online educational platforms, namely `Vuihoc.vn`, `Vietjack.com`, and `Yourhomework.net`. These platforms host an extensive repository of educational materials, including lectures, practice exercises, and resources that frequently incorporate charts and tables integral to pedagogical practices. The collected images encompass a diverse array of chart types, such as bar charts, line charts, and pie charts, as well as various table formats, including numerical tables, comparative tables, and analytical tables, spanning subjects such as mathematics, economics, and data analysis.

The selection of these platforms was driven by their use of data sourced exclusively from Vietnam's official statistical yearbooks, which are renowned for their accuracy and authority. Compiled by Vietnam's General Statistics Office and other authoritative bodies, these yearbooks provide meticulously verified and standardized data, serving as a cornerstone for research activities in Vietnam. By grounding their charts and tables in such reputable sources, these platforms ensure the precision and trustworthiness of their visual data. This robust foundation of credible data positions the **ChartTableViVQA** dataset as a reliable resource for developing VQA models tailored to the Vietnamese context, advancing research in automated question answering for visual data.

3.2   Question-Answer Generation

Following the collection of chart and table images, the subsequent phase involves the automated generation of questions. We employed the GPT language model to create Visual Question Answering (VQA) questions based on each collected image. These questions are designed to evaluate the ability to comprehend and analyze information presented in charts and tables. Below, we outline the four primary question types generated, along with detailed explanations for each:

– **Direct Data Retrieval (Lookup):** These questions require extracting a specific value from a chart or table without performing any computations. Such questions are straightforward, relying on the model's ability to identify a value at a designated row or column in a table or a specific data point in a chart.
– **Basic Arithmetic Operations (Aggregation):** These questions mandate the model to perform fundamental mathematical operations, such as calculating sums, averages, or differences, using values from a chart or table. The

objective is to assess the model's computational proficiency with numerical data. For instance, a question may require subtracting values from two different years to derive a result.

– **Comparison:** These questions task the model with comparing values within a chart or table to determine which is greater, lesser, or to quantify changes between them. This question type evaluates the model's understanding of relationships among values and its ability to draw conclusions about their differences. For example, a question may involve comparing area values from two distinct years and computing their difference.

– **Trend Analysis:** These questions require the model to analyze data changes over time or across groups to identify patterns of growth or decline. Trend-focused questions demand that the model recognize variations in data across years or regions and provide insights into increases or decreases. For instance, a question may ask the model to identify the overall trend in data over multiple years and comment on its evolution.

3.3   Instruction Generation

After generating question-answer pairs from chart and table images, we proceed to construct a detailed set of instructions for each question. The purpose of this phase is to clearly outline the necessary steps to extract information from charts or tables, aiding the model in understanding the context and the reasoning process required to produce accurate answers. These instructions are designed to support both the training and inference processes of the model.

The instructions are constructed according to the different types of questions, which include:

– **Direct Data Retrieval (Lookup):** These instructions specify how to identify a particular value from the chart or table, based on certain indicators or time points.

– **Basic Arithmetic Operations (Aggregation):** These instructions describe how to perform simple arithmetic operations, such as addition, subtraction, averages, or percentage changes between values.

– **Comparison:** These instructions guide the model to compare values at different time points or across different entities to determine differences or identify trends.

– **Trend Analysis:** These instructions require the model to analyze data over time or across groups, identifying patterns of increase or decrease.

These instructions not only help the model grasp the reasoning process behind each answer but also play a crucial role in clarifying and making the model's decisions more interpretable in multimodal VQA tasks. They support the model in understanding the steps of computation and logic, while ensuring accuracy and consistency in the answer generation process.

By providing clear instructions, the dataset not only offers answers but also helps shape the reasoning process, opening up possibilities for integration with

Table 1: *Examples of Questions, Answers, and Instructions in Vietnamese and English for the ChartTableViVQA Dataset*

Questions, Answers, and Instructions

Question: Diện tích trồng chuối vào năm 2021 là bao nhiêu nghìn ha? [English: What was the area of banana cultivation in Vietnam in 2021 in thousand hectares?]
Answer: 154.2 nghìn ha [English: 154.2 thousand hectares]
Instruction: Tìm giá trị của 'Diện tích (nghìn ha)' tại Năm '2021' là 154.2. [English: Find the value of 'Area (thousand hectares)' for the year '2021' which is 154.2.]

Question: Chênh lệch sản lượng nuôi trồng thuỷ sản giữa năm 2013 và năm 2005 là bao nhiêu nghìn tấn? [English: What is the difference in aquaculture production between 2013 and 2005 in thousand tonnes?]
Answer: 1737.9 nghìn tấn [English: 1737.9 thousand tonnes]
Instruction: Lấy giá trị hàng 'Nuôi trồng' tại cột '2013' là 3215.9 và trừ giá trị tại cột '2005' là 1478.0: 3215.9 - 1478.0 = 1737.9. [English: Take the value in the 'Aquaculture' row for the '2013' column which is 3215.9 and subtract the value in the '2005' column which is 1478.0: 3215.9 - 1478.0 = 1737.9.]

Question: Năm 2013, diện tích cây công nghiệp lâu năm ở cả nước có cao hơn năm 2005 không? [English: In 2013, was the area of perennial industrial crops in Vietnam higher than in 2005?]
Answer: Cao hơn 501.3 nghìn ha [English: Higher by 501.3 thousand hectares]
Instruction: So sánh giá trị tại hàng 'Cây công nghiệp lâu', cột 'Cả nước (nghìn ha)' năm 2013 là 2134.9 và năm 2005 là 1633.6: 2134.9 > 1633.6, 2134.9 - 1633.6 = 501.3 (Cao hơn). [English: Compare the value in the 'Perennial industrial crops' row, 'Nationwide (thousand hectares)' column for 2013 which is 2134.9 and for 2005 which is 1633.6: 2134.9 > 1633.6, 2134.9 - 1633.6 = 501.3 (Higher).]

Question: Năm 2021, GDP của Bru-nây tăng bao nhiêu phần trăm so với năm 2020? [English: In 2021, by what percentage did Brunei's GDP increase compared to 2020?]
Answer: Tăng 22.73% [English: Increased by 22.73%]
Instruction: Tính phần trăm thay đổi bằng cách lấy (giá trị của cột 'Bru-nây (nghìn USD)' tại Năm '2021' là 32.4 trừ giá trị tại Năm '2020' là 26.4) chia cho giá trị tại Năm '2020' là 26.4, rồi nhân với 100: ((32.4 - 26.4) / 26.4) * 100 = 22.73 (Tăng). [English: Calculate the percentage change by taking (the value in the 'Brunei (thousand USD)' column for the year '2021' which is 32.4 minus the value for the year '2020' which is 26.4) divided by the value for the year '2020' which is 26.4, then multiply by 100: ((32.4 - 26.4) / 26.4) * 100 = 22.73 (Increased).]

instruction-based models or serving as training data for explainable AI tasks. The development of detailed instructions for each question also ensures consistency and quality control of the dataset, while assisting in error analysis during later stages by clarifying whether the model's mistakes arise from misunderstanding the question, incorrect computations, or faulty data extraction from the charts.

3.4   Data Format for Charts and Tables

In the context of Visual Question Answering (VQA) for charts and tables, a standardized data format is essential to ensure consistency and usability across the dataset. This format allows for seamless integration of image data, related metadata, and question-answer pairs, supporting the training and evaluation of VQA models. We have developed a JSON-based structure to systematically encapsulate all critical information for each chart and table image.

For chart images, the JSON structure includes the following main fields:

- **document_id**: A unique identifier.
- **image_info**:
- **image_name**: The image file name.
- **chart_type**: The chart type (e.g., pie chart, bar chart).
- **metadata**: Information such as title and axis labels.
- **caption**: A brief description of the chart's content.
- **data**: Chart data in a tabular format, typically a list of lists.
- **data_text**: A textual summary of the data.
- **qa**: A list of question-answer pairs, each including:
- **qa_id**: The question-answer pair identifier.
- **question**: The question.
- **answer**: The answer.
- **instruction**: Guidance to derive the answer from the chart.

Similarly, for table images, the JSON structure is designed to effectively represent tabular data:

- **document_id**: A unique identifier.
- **image_info**:
- **image_name**: The image file name.
- **caption**: A description of the table content.
- **columns**: A list of column names.
- **rows**: A list of dictionaries, each representing a row with keys corresponding to column names.
- **metadata**: Additional information such as measurement units.
- **qa**: A list of question-answer pairs and instruction, similar to those for charts.

Detailed examples of the JSON structures for both charts and tables are provided in the Appendix. These examples illustrate the application of the format and how data is organized to support the diverse question types described in the "Question-Answer Generation"section. By establishing this standardized data format, we aim to facilitate future research in chart and table VQA, enabling the development and benchmarking of new models, exploration of advanced question types, or expansion of the dataset with additional images and questions.

3.5   Annotation and Validation

Following the automated generation of questions and answers by the GPT model, the annotation and validation phase is conducted to ensure the quality of the **ChartTableViVQA** dataset. This phase comprises three steps designed to identify and rectify errors in question-answer pairs, categorized into four primary error types: unclear questions, misspellings, incorrect answers, and answers with missing or redundant information. The three steps are described as follows:

1. **Semantic Validation**: Annotators assess the clarity, coherence, and naturalness of each question to ensure it is unambiguous and comprehensible, thereby enhancing its suitability for VQA tasks.
2. **Accuracy Verification**: Annotators verify that answers accurately correspond to the posed questions and that any calculations, if applicable, are performed correctly, ensuring data integrity.
3. **Cross-Checking**: Annotators cross-review each other's question-answer pairs. Upon identifying errors, they engage in discussions to reach a consensus and make corrections, ensuring consistency and quality across the dataset.

This rigorous validation process ensures that the **ChartTableViVQA** dataset meets high standards, ready for training and evaluating VQA models in the Vietnamese context.

3.6   Dataset Analysis

Figure 2 illustrates the distribution of question types within the **ChartVQA** and **TableVQA** datasets, offering valuable insights into the nature of visual and tabular question-answering tasks. In the ChartVQA dataset, nearly half of the questions fall under the "Lookup"category, with "Comparison"and "Trend Analysis"following as significant components. Similarly, the TableVQA dataset is dominated by "Lookup"questions, but it also features a slightly higher proportion of "Aggregation"and "Trend Analysis"inquiries. These variations underscore the distinct characteristics and complexities of interpreting charts versus tables.

To enhance clarity, rare categories have been excluded from the analysis.

4   Methodology

To evaluate the capacity of multimodal language models in interpreting Vietnamese chart and table data, we design an end-to-end benchmarking pipeline centered on the newly constructed **ChartTableViVQA** dataset. Our methodology consists of two core components: the selection of a strong vision-language baseline model and the implementation of a comprehensive evaluation framework to assess generated answers across diverse question types.
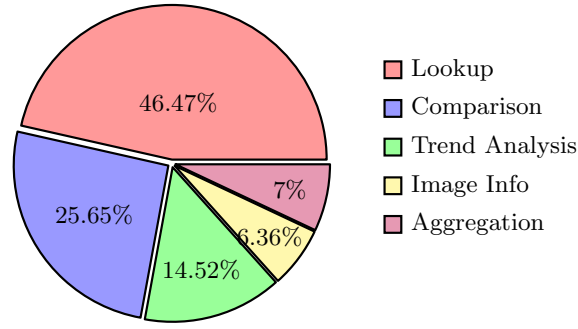
(a) Question Type of ChartVQA
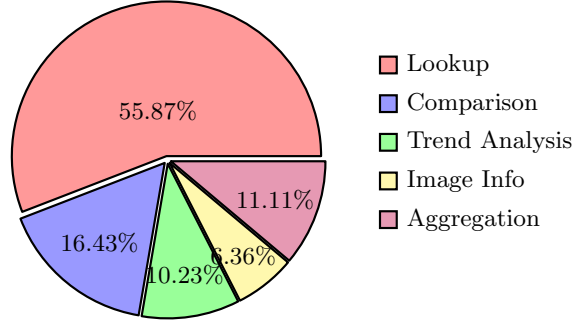
(b) Question Type of TableVQA

Figure 2: *The distribution of question types in the ChartVQA and TableVQA datasets (Other merged into Aggregation).*

4.1   Baseline Model

We adopt the **Qwen2.5-VL-7B-Instruct** model as our baseline, a state-of-the-art vision-language model with 7 billion parameters. This large-scale model is designed to handle complex multimodal tasks, requiring fine-grained reasoning over both visual and textual information. Its architecture combines a high-capacity visual encoder with an instruction-tuned language decoder, enabling robust performance on educational Visual Question Answering (VQA) tasks that involve Vietnamese charts and tables.

For inference, the model processes each input by using a set of dynamically generated image patches and a natural language question embedded within a carefully crafted prompt. Each input consists of a chart or table image, an instruction, and the question. This setup ensures that the model is grounded in visual evidence while being guided to generate concise and contextually appropriate answers. These answers align with the measurement units and temporal references present in the source data, demonstrating the model's ability to interpret and respond to complex Visual Question Answering (VQA) tasks.

4.2   Evaluation Framework

To assess the performance of the **Qwen2.5-VL-7B-Instruct** model on the ChartTableViVQA dataset, we utilize a standard evaluation framework consisting of five commonly used metrics. These metrics are widely applied in tasks such as machine translation, text summarization, and VQA to quantify the quality of generated responses with respect to human-labeled reference answers.

- **BLEU [11]**: BLEU is a widely-used n-gram overlap metric that measures the precision of the generated text against the reference. The score is calculated using BLEU-1 through BLEU-4, corresponding to unigrams through 4-grams. We apply uniform weights across the n-grams and use smoothing techniques to handle cases where shorter answers may result in zero counts for specific n-grams. BLEU captures lexical accuracy and is effective for evaluating short, structured outputs typical in factual VQA tasks.
- **METEOR [12]**: METEOR computes the harmonic mean of unigram precision and recall, with an emphasis on recall. It includes advanced matching techniques such as stemming and synonymy, which improve its ability to capture semantic similarity. In this work, we compute METEOR using exact word matching due to the current lack of support for advanced features in Vietnamese. This metric is useful for evaluating answers that may vary in phrasing but are semantically similar.
- **ROUGE [13]**: ROUGE is a recall-based metric, primarily used in automatic summarization tasks, which measures the overlap of unigrams (ROUGE-1) and bigrams (ROUGE-2) between the predicted and reference answers. These scores are computed using stemming and longest common subsequence (LCS) matching, which helps capture relevant content in the predicted answers even when the word order differs from the reference.
- **BERTScore [14]**: BERTScore computes semantic similarity by comparing contextual embeddings of tokens using a pretrained BERT model. We use the multilingual BERT model `bert-base-multilingual-cased` with a Vietnamese language setting to calculate the similarity between the generated and reference answers. This metric is particularly effective in capturing nuanced semantic differences and recognizing paraphrased or rephrased answers.

All metrics are computed for each question-answer pair and then averaged over the entire dataset to provide overall performance scores. The inclusion of both lexical (BLEU, ROUGE, METEOR) and semantic (BERTScore) metrics ensures that the evaluation captures both precision in form and adequacy in content. In the case of errors during metric computation (e.g., due to malformed text), the score for that instance is set to zero to ensure consistency across the evaluation process.

These metrics provide a comprehensive view of the model's performance, allowing for detailed analysis and comparison across different question categories and answering strategies.

5   Experimental Results

5.1   Experimental Preparation

The model inference and training processes were carried out in the **Kaggle** environment, which provides access to **two NVIDIA Tesla T4 GPUs**. These GPUs offer substantial computational power for efficiently handling the multimodal inference tasks required in this study. The hardware setup enables efficient parallel processing, allowing the model to process large volumes of data, including high-resolution chart and table images, in a timely manner.

By utilizing the GPU-accelerated setup, we were able to ensure smooth execution of the complex **Qwen2.5-VL-7B-Instruct** model, enabling the processing of dynamic image patches and the associated natural language questions. This environment ensures high-performance, resource-optimized execution of the model while maintaining accuracy and computational efficiency.

5.2   Evaluation Results

The performance of the **Qwen2.5-VL-7B-Instruct** baseline model was assessed on the **ChartTableViVQA** dataset, which includes both the chart and table sections. The results are summarized in Table 2, where the performance is expressed as percentages for each metric.

Metric	BERTScore	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	METEOR
ChartVQA (%)	76.70	15.50	10.80	8.10	6.50	39.80	26.80	18.70
TableVQA (%)	82.20	29.10	23.10	17.40	13.40	55.00	46.30	34.90
Combined Average Metrics (%)	79.45	22.30	16.95	12.75	9.95	47.90	36.55	26.80

Table 2: *Performance metrics for the Qwen2.5-VL-7B-Instruct model on ChartVQA and TableVQA datasets, including average values for both datasets.*

The model's performance shows variability across different evaluation metrics, highlighting its strengths and weaknesses in multimodal tasks. Specifically, the **BERTScore**, which measures the semantic similarity between the generated answers and the reference answers, achieves 76.7% for **ChartVQA** and 82.2% for **TableVQA**. These results suggest a reasonable alignment with the reference answers, with a slight improvement observed on the **TableVQA** dataset. This improvement can be attributed to the more structured nature of tabular data, which is easier for the model to process compared to the more complex layouts found in chart data.

Regarding the **BLEU score**, which assesses the n-gram precision between the predicted answers and the reference answers, the results are relatively modest. The **BLEU-1** score is 15.5% for **ChartVQA** and 29.1% for **TableVQA**, while the **BLEU-4** score decreases to 6.5% and 13.4%, respectively. This suggests that while the model captures some exact matches, it struggles with longer sequences, particularly when dealing with chart data. These results imply that

the model's ability to generate longer and more complex sequences could be further improved.

In contrast, the **ROUGE scores**, which focus on recall by measuring the overlap of n-grams between the generated answers and the reference answers, show more favorable results. The **ROUGE-1** score is 39.8% for **ChartVQA** and 55.0% for **TableVQA**, while the **ROUGE-2** score is 26.8% and 46.3%, respectively. These values suggest that the model is more successful in capturing the key content, especially in the **TableVQA** tasks, where the structured nature of the data facilitates more efficient information extraction compared to less structured charts.

The **METEOR score**, which evaluates the semantic adequacy of the generated answers, is 18.7% for **ChartVQA** and 34.9% for **TableVQA**. These results reflect the model's moderate ability to generate contextually appropriate responses, with **TableVQA** showing better performance due to the clearer structure of the tabular data.

Overall, the stable performance of the **Qwen2.5-VL-7B-Instruct** model on the **ChartTableViVQA** dataset can be attributed to its integrated **vision-language (VL)** architecture, which effectively combines pretrained image encoders with a language model. This multimodal design enables the model to understand complex visual elements such as charts and tables while integrating them with text queries in a coherent manner. The model's training on various types of image data during pretraining likely enhanced its ability to extract meaningful patterns from the structured layouts of **ChartTableViVQA**, particularly for **TableVQA**, where organized data structures assist in feature extraction more effectively compared to the less structured charts. However, the limitations in n-gram accuracy indicate that the model's ability to generate sequences still needs improvement, especially when handling less structured inputs such as charts.

5.3   Error Analysis

Through the analysis of predicted answers from the ChartTableViVQA dataset, we identified three main error categories encountered by the Qwen2.5-VL-7B-Instruct model during the answer generation process, as shown in Table 3.

- **Error in trend or contradiction with the correct answer**: One common issue is the model's inability to accurately determine the trend (increase/decrease) of values or to draw conclusions consistent with the ground truth. Although the model may provide approximate numerical data, it often fails to grasp the change or trend in the question, leading to results that contradict the truth. This may stem from the model's limited ability to recognize relationships between values in chart/table data.
- **Error in mismatch between answer phrasing and ground truth**: The model may provide correct numerical answers, but the phrasing does not align with the ground truth. This results in inconsistency between the model's generated answer and the standard answer. These errors are primarily related to the model's failure to correctly interpret the semantic requirements

Table 3: *Error Analysis for Questions and Answers in Vietnamese and English for the ChartTableViVQA Dataset*

Questions and Answers
Question: Tỷ lệ lao động khu vực nông - lâm - ngư nghiệp năm 2005 có cao hơn dịch vụ không? [English: Is the labor ratio in the agriculture, forestry, and fishery sector in 2005 higher than that in the service sector?] **Answer**: Có, cao hơn 32.8%. [English: Yes, higher by 32.8%.] **Prediction**: Không, tỷ lệ lao động khu vực nông - lâm - ngư nghiệp năm 2005 là 57,3%, còn tỷ lệ lao động khu vực dịch vụ năm 2005 là 18,2%. [English: No, the labor ratio in the agriculture, forestry, and fishery sector in 2005 was 57.3%, while the labor ratio in the service sector in 2005 was 18.2%.]
Question: Giá trị xuất khẩu của Xin-ga-po tăng hay giảm bao nhiêu phần trăm từ năm 2010 đến 2013? [English: Did the export value of Singapore increase or decrease by how many percent from 2010 to 2013?] **Answer**: Tăng 28.39%. [English: Increased by 28.39%.] **Prediction**: Giá trị xuất khẩu của Xin-ga-po tăng 6 phần trăm từ năm 2010 đến 2013. [English: The export value of Singapore increased by 6 percent from 2010 to 2013.]
Question: Chênh lệch xuất khẩu giữa năm 2014 và 2012 là bao nhiêu? [English: What is the export difference between 2014 and 2012?] **Answer**: 15.2 tỷ đô la Mỹ. [English: 15.2 billion US dollars.] **Prediction**: -15,2 tỷ đô la Mỹ. [English: -15.2 billion US dollars.]
Question: Nhiệt độ tháng V tăng hay giảm so với tháng IV bao nhiêu? [English: Did the temperature in May increase or decrease compared to April by how much?] **Answer**: Tăng 4.0°C. [English: Increased by 4.0°C.] **Prediction**: 29.3°C - 25.3°C = 4°C tăng. [English: 29.3°C - 25.3°C = 4°C increased.]
Question: Sản lượng ngô của Thái Lan thay đổi như thế nào từ năm 2019 đến năm 2020? [English: How did the corn production of Thailand change from 2019 to 2020?] **Answer**: Tăng 292.17 nghìn tấn. [English: Increased by 292.17 thousand tons.] **Prediction**: Sản lượng ngô của Thái Lan giảm từ 4 703,0 nghìn tấn vào năm 2019 xuống còn 4 995,17 nghìn tấn vào năm 2020. [English: The corn production of Thailand decreased from 4,703.0 thousand tons in 2019 to 4,995.17 thousand tons in 2020.]
Question: Sản lượng ngô của Việt Nam thay đổi như thế nào từ năm 2019 đến năm 2020? [English: How did the corn production of Vietnam change from 2019 to 2020?] **Answer**: Giảm 326.38 nghìn tấn. [English: Decreased by 326.38 thousand tons.] **Prediction**: Sản lượng ngô của Việt Nam giảm từ 4 732,2 nghìn tấn vào năm 2019 xuống còn 4 405,82 nghìn tấn vào năm 2020. [English: The corn production of Vietnam decreased from 4,732.2 thousand tons in 2019 to 4,405.82 thousand tons in 2020.]

of the question or to convey information in a manner consistent with the original answer.

– **Error in quantitative accuracy**: A frequent error is the model's inaccuracy in calculating or converting quantitative values from chart or table data into answers. The model may provide incorrect numerical values or fail to accurately compute percentages, figures, or differences between years. This could be due to inaccuracies in the model's process of converting image-based data into text or errors in its computational logic.

These errors indicate that the Qwen2.5-VL-7B-Instruct model is not yet fully accurate in understanding the context of questions and converting image data into content-accurate answers. Issues related to mismatches between quantitative values and phrasing are areas that need improvement in future model versions.

6 Conclusion and Future Work

In this study, we introduced the **ChartTableViVQA** dataset, a pioneering resource for Vietnamese Visual Question Answering (VQA) tasks involving chart and table images. This dataset comprises 395 chart images and 474 table images sourced from reputable Vietnamese educational websites, accompanied by 13,069 self-generated question-answer pairs. The diversity of question categories, including calculations, comparisons, and trend analysis, allows for a comprehensive evaluation of models' ability to interpret both visual content and semantic information in multimodal tasks.

These observations underscore the model's potential in interpreting visual elements and aligning them with textual queries. However, the lower performance in precise n-gram matching suggests that the model might struggle with tasks requiring exact, word-for-word correspondence. The performance differences between **ChartVQA** and **TableVQA** can likely be attributed to the more organized nature of tables, which facilitate easier information extraction compared to the more varied and complex structures of charts. Despite these promising results, there are still areas for improvement, particularly in sequence generation and handling of linguistic subtleties, to further enhance the model's effectiveness across both chart and table-based tasks.

The **Qwen2.5-VL-7B-Instruct** model was evaluated on this dataset, achieving reasonable performance, with improvements observed on the **TableVQA** subset due to the more structured nature of tabular data. The analysis highlighted the model's ability to capture semantic similarities, although challenges remain in handling long sequences and ensuring the accurate generation of quantitative answers. This work provides valuable insights into the strengths and weaknesses of current models, serving as a foundation for advancing VQA systems tailored to the Vietnamese context.

In the future, we aim to extend the **ChartTableViVQA** dataset by incorporating additional question types and more diverse data sources, further enriching its applicability. A key direction for future work is enhancing model generalization, particularly through the inclusion of paraphrased answers, which may improve the model's robustness in handling varied phrasing. Furthermore, additional

research will explore the potential for refining the model's ability to handle less structured data, such as charts, while continuing to improve quantitative accuracy and semantic coherence in answers.

This study sets the groundwork for the development of Vietnamese-language VQA systems and lays the foundation for future research in multimodal AI models. The **ChartTableViVQA** dataset serves as an important resource for the advancement of VQA models in Vietnamese, opening up opportunities for further scalability across diverse domains and languages.

A   Examples of JSON Data Structures

A.1   Describing Chart Example

The following structure exemplifies the data format for a chart image, specifically a pie chart representing the proportion of international tourists arriving in Vietnam by mode of transportation for the years 2010, 2012, and 2015. The structure includes the chart's metadata, data, textual summary, and associated question-answer pairs.

```
{
  "document_id": "0870",
  "image_info": {
    "image_name": "0870.PNG",
    "chart_type": "pie_chart",
    "metadata": {
      "title": "BIỂU ĐỒ CƠ CẤU KHÁCH DU LỊCH QUỐC
      TẾ ĐẾN VIỆT NAM PHÂN THEO PHƯƠNG TIỆN ĐẾN",
      "x_axis_label": "Năm",
      "y_axis_label": "Tỷ lệ (%)"
    },
    "caption": "Biểu đồ thể hiện cơ cấu khách du lịch quốc tế đến Việt Nam
    phân theo phương tiện đến trong các năm 2010, 2012 và 2015."
  },
  "data": [
    ["Năm", "Phương tiện", "Tỷ lệ (%)"],
    ["2010", "Đường hàng không", 80.4],
    ["2010", "Đường thủy", 18.6],
    ["2010", "Đường bộ", 1.0],
    ["2012", "Đường hàng không", 81.4],
    ["2012", "Đường thủy", 14.4],
    ["2012", "Đường bộ", 4.2],
    ["2015", "Đường hàng không", 79.0],
    ["2015", "Đường thủy", 19.0],
    ["2015", "Đường bộ", 2.0]
  ],
  "data_text": "Năm 2010: Đường hàng không 80.4%, Đường thủy 18.6%, Đường bộ 1.0%.
Năm 2012: Đường hàng không 81.4%, Đường thủy 14.4%, Đường bộ 4.2%.
Năm 2015: Đường hàng không 79.0%, Đường thủy 19.0%, Đường bộ 2.0%.",
  "qa": [
    {
      "qa_id": "q1",
      "question": "Biểu đồ trên thể hiện nội dung gì?",
      "answer": "Biểu đồ thể hiện cơ cấu khách du lịch
      quốc tế đến Việt Nam phân theo phương tiện đến trong các năm 2010,
      2012 và 2015.",
      "instruction": "BIỂU ĐỒ CƠ CẤU KHÁCH DU LỊCH QUỐC TẾ ĐẾN
      VIỆT NAM PHÂN THEO PHƯƠNG TIỆN ĐẾN"
    }
    ...
```

```
  ]
}
```

Below is the English version of the same structure.

```
{
  "document_id": "0870",
  "image_info": {
    "image_name": "0870.PNG",
    "chart_type": "pie_chart",
    "metadata": {
      "title": "CHART OF THE STRUCTURE OF INTERNATIONAL TOURISTS ARRIVING
      IN VIETNAM BY MODE OF TRANSPORTATION",
      "x_axis_label": "Year",
      "y_axis_label": "Percentage (%)"
    },
    "caption": "The chart shows the structure of international tourists
    arriving in Vietnam by mode of transportation for the years 2010, 2012, and 2015."
  },
  "data": [
    ["Year", "Mode of Transportation", "Percentage (%)"],
    ["2010", "Air", 80.4],
    ["2010", "Sea", 18.6],
    ["2010", "Land", 1.0],
    ["2012", "Air", 81.4],
    ["2012", "Sea", 14.4],
    ["2012", "Land", 4.2],
    ["2015", "Air", 79.0],
    ["2015", "Sea", 19.0],
    ["2015", "Land", 2.0]
  ],
  "data_text": "Year 2010: Air 80.4%, Sea 18.6%, Land 1.0%.
  Year 2012: Air 81.4%, Sea 14.4%, Land 4.2%.
  Year 2015: Air 79.0%, Sea 19.0%, Land 2.0%.",
  "qa": [
    {
      "qa_id": "q1",
      "question": "What does the chart represent?",
      "answer": "The chart represents the structure of international tourists
      arriving in Vietnam by mode of transportation for the years 2010,
      2012, and 2015.",
      "instruction": "CHART OF THE STRUCTURE OF INTERNATIONAL TOURISTS
      ARRIVING IN VIETNAM BY MODE OF TRANSPORTATION"
    }
    ...
  ]
}
```

A.2   Describing Table Example

The following structure represents a table detailing the export values of selected commodities (pepper, coffee, and rubber) in Vietnam from 2015 to 2021, including metadata and question-answer pairs based on the provided image.

```
{
  "document_id": "0405",
  "image_info": {
    "image_name": "0405.jpg",
    "caption": "Trị giá một số mặt hàng xuất khẩu của nước ta, giai đoạn 2015 - 2021"
  },
  "columns": ["Năm", "2015", "2019", "2020", "2021"],
  "rows": [
    { "Năm": "Hạt tiêu", "2015": 1259.9, "2019": 714.4, "2020": 660.6, "2021": 937.8 },
    { "Năm": "Cà phê", "2015": 2671.0, "2019": 2863.8, "2020": 2742.0, "2021": 3072.6 },
    { "Năm": "Cao su", "2015": 1531.5, "2019": 2300.7, "2020": 2384.4, "2021": 3278.3 }
  ],
  "metadata": {
    "Đơn vị": "Giá trị (Triệu USD)"
  },
  "qa": [
    { "qa_id": "q1", "question": "Bảng trên thể hiện nội dung gì?",
    "answer": "Bảng trên thể hiện trị giá một số mặt hàng xuất khẩu của nước ta,
    giai đoạn 2015 - 2021.", "instruction": "Cho bảng số liệu về trị giá một
    số mặt hàng xuất khẩu của nước ta, giai đoạn 2015 - 2021." },
    { "qa_id": "q2", "question": "Giá trị Cao su năm 2021 là bao nhiêu?",
    "answer": "3278.3 triệu USD.", "instruction": "Tìm giá trị tại cột 2021,
    hàng Cao su là 3278.3 triệu USD." }
  ]
}
```

Below is the English version of the same structure.

```
{
  "document_id": "0405",
  "image_info": {
    "image_name": "0405.jpg",
    "caption": "Export values of selected commodities in Vietnam, period 2015 - 2021"
  },
  "columns": ["Year", "2015", "2019", "2020", "2021"],
  "rows": [
    { "Year": "Pepper", "2015": 1259.9, "2019": 714.4, "2020": 660.6, "2021": 937.8 },
    { "Year": "Coffee", "2015": 2671.0, "2019": 2863.8, "2020": 2742.0, "2021": 3072.6 },
    { "Year": "Rubber", "2015": 1531.5, "2019": 2300.7, "2020": 2384.4, "2021": 3278.3 }
  ],
  "metadata": {
    "Unit": "Value (Million USD)"
  },
  "qa": [
```

```
   { "qa_id": "q1", "question": "What does the table represent?",
   "answer": "The table represents the export values of selected
   commodities in Vietnam, period 2015 - 2021.",
   "instruction": "Provide table data on the export values of selected
   commodities in Vietnam, period 2015 - 2021." },
   { "qa_id": "q2", "question": "What is the export value of rubber in 2021?",
   "answer": "3278.3 million USD.",
   "instruction": "Find the value in the 2021 column, Rubber row,
   which is 3278.3 million USD." }
  ]
}
```

B   Examples of Image

B.1   Chart Image Example
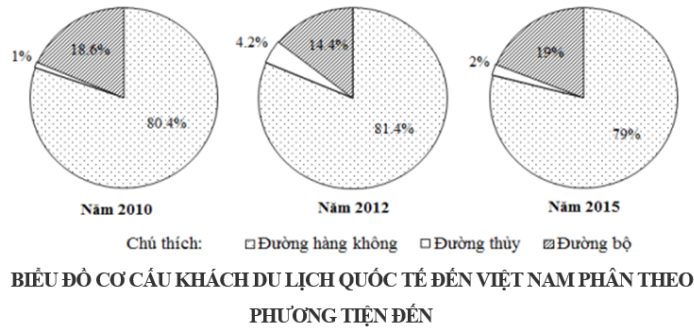
Figure 3: *Pie chart showing the proportion of international tourists arriving in Vietnam by mode of transportation (2010, 2012, 2015) document ID 0870.*

B.2   Table Image Example

Figure 4: *Table showing the export values of selected commodities in Vietnam (2015-2021) document ID 0405.*

References

1. Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024.
2. Hui Shi, Yusheng Xie, Luis Goncalves, Sicun Gao, and Jishen Zhao. Wikidt: Visual-based table recognition and question answering dataset. In *International Conference on Document Analysis and Recognition*, pages 406–437. Springer, 2024.
3. Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. Multimodal table understanding. *arXiv preprint arXiv:2406.08100*, 2024.
4. Brandon Smock, Rohith Pesala, and Robin Abraham. Pubtables-1m: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4634–4642, 2022.
5. Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. Tablebank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, 2020.
6. Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*, 2021.
7. Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.
8. Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. Leaf-qa: Locate, encode & attend for figure question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3512–3521, 2020.
9. Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. Chartqa-x: Generating explanations for charts. *arXiv preprint arXiv:2504.13275*, 2025.
10. Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
11. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.
12. Satanjeev Banerjee and Alon Lavie. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. *Proceedings of ACL-WMT*, pages 65–72, 2004.
13. Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.
14. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.