ChartTableViVQA: A Vietnamese Benchmark Dataset for Visual Question Answering with Reasoning Instruction over Charts and Tables Images

Quang Nhat Truong[1,2], **Thang Quoc Pham**[1,2], **Son T. Luu**[1,2]*
[1]Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{22521207,23521431}@gm.uit.edu.vn, sonlt@uit.edu.vn

Abstract

Visual Question Answering (VQA) over charts and tables requires joint reasoning over language, visual structure, and numerical information, yet no standard benchmark currently exists for Vietnamese. We introduce **ChartTableViVQA**, a Vietnamese VQA dataset with 13,003 question–answer–instruction triplets over 868 chart and table images. Questions cover multiple reasoning types, ranging from simple value lookup to multi-step aggregation, comparison, and trend analysis, and are accompanied by concise instructions that explicitly guide the reasoning process. The dataset is built via a semi-automatic human-in-the-loop pipeline, combining large language model generation with manual verification by Vietnamese annotators. We evaluate several vision-language models in zero-shot and instruction-based settings. Results show strong performance on simple questions but clear weaknesses in complex reasoning, particularly on charts, while instruction-based prompting consistently improves accuracy.

1 Introduction

Visual Question Answering (VQA) over structured visuals such as charts and tables is a challenging testbed for multimodal reasoning. Unlike natural images, charts and tables encode information through axes, legends, cell structures, and numerical values, requiring models to jointly interpret language, layout, and quantitative relationships. Such capabilities are essential in real-world domains, including education, official statistics, and data-driven reporting.

Recent chart and table VQA benchmarks have driven rapid progress by covering diverse visual encodings and question types, ranging from direct value lookup to aggregation and comparison. However, existing datasets are almost exclusively available in English and assume high-resource settings. For low-resource languages such as Vietnamese, there is still no standardized benchmark for structured visual reasoning, making it difficult to systematically evaluate current vision-language models (VLMs) and analyze their limitations.

To address this gap, we introduce **ChartTableViVQA**, a Vietnamese VQA dataset for chart and table images. The dataset contains 13,003 question–answer–instruction triplets grounded in 868 images collected from official statistical sources. Each question is annotated with one of four reasoning types—value lookup, aggregation, comparison, and trend analysis—and assigned to three difficulty levels. In addition, each QA pair is accompanied by a concise instruction describing how to locate relevant values in the visual input and perform the required reasoning steps.

ChartTableViVQA is constructed through a semi-automatic, human-in-the-loop pipeline. A large language model is first used to generate candidate questions, answers, and instructions, which are then reviewed and corrected by Vietnamese annotators to ensure semantic consistency with the visual content and numerical correctness. This design balances annotation quality with scalability in a low-resource language setting.

Using **ChartTableViVQA**, we evaluate several recent VLMs, including both multilingual and Vietnamese-centric models, under two inference settings: zero-shot and instruction-based. Our evaluation focuses on three research questions:

- **RQ1**: How well do current VLMs handle Vietnamese chart and table VQA across different reasoning types and difficulty levels?
- **RQ2**: How does performance differ between chart-based and table-based questions?
- **RQ3**: To what extent do explicit instructions improve performance on multi-step reasoning

*Corresponding author

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 99–123
November 6-10, 2026 ©2026 Association for Computational Linguistics

tasks?

Our results show that even strong models perform reliably on simple lookup questions but struggle with multi-step reasoning, particularly on charts. Providing explicit instructions leads to substantial performance improvements, especially for aggregation, comparison, and trend analysis, but does not fully close the gap on the most challenging cases.

Our contributions are summarized as follows:

- We introduce **ChartTableViVQA**, the first Vietnamese VQA benchmark for structured reasoning over charts and tables, comprising 13,003 triplets across four reasoning types and three difficulty levels.
- We propose an LLM-assisted, human-in-the-loop data construction pipeline suitable for low-resource languages.
- We conduct a systematic evaluation of multilingual and Vietnamese-centric VLMs under zero-shot and instruction-based settings.
- We provide an error analysis highlighting common failure modes in numerical and comparative reasoning.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 introduces the ChartTableViVQA dataset and details the data collection process, question–answer and instruction generation, data format, and analysis. Section 4 outlines the experimental setup. Section 5 presents the evaluation results and error analysis. Section 6 concludes with discussions and future directions.

2 Related Work

Vision-based table question answering (table VQA) focuses on answering natural language questions directly from table images, such as scanned documents or screenshots, without relying on underlying table markup. This setting is common in real-world document understanding scenarios. Additionally, Chart question answering (chart VQA) aims to answer questions about visualized data, such as bar, line, and pie charts. Compared to table VQA, chart VQA requires interpreting graphical encodings in addition to textual and numerical information.

Several datasets that support table and chart understanding from images are constructed. For example, TableVQA-Bench (Kim et al., 2024) introduces QA pairs over rendered table images across multiple domains, WikiDT (Shi et al., 2024) provides web-style document images with table QA and intermediate annotations for retrieval and structure analysis, ChartQA (Masry et al., 2022) is a widely adopted benchmark emphasizing visual and logical reasoning over charts, and ChartQA-X (Hegde et al., 2025) extends chart VQA by introducing natural language explanations to justify model predictions. Despite these advances, most existing table VQA datasets are constructed in English and focus on question–answer supervision. Moreover, the supervision is typically limited to final answers or free-form explanations, without concise step-by-step instructions that guide reasoning from visual evidence.

Recent frontier vision–language models have demonstrated strong performance on chart and table VQA tasks. Closed-source models such as GPT-4o (OpenAI et al., 2024) and Gemini-2.5-Flash (Comanici et al., 2025) have established new state-of-the-art results in multimodal understanding and complex visual reasoning. Open-source models, including Qwen-VL (Bai et al., 2023), InternVL (Chen et al., 2024), and Llama-3.2-Vision (Grattafiori et al., 2024), have significantly scaled vision–language pretraining for generic multimodal reasoning. In addition, Vietnamese-focused open-source models such as LaVy (Tran and Thanh, 2024) and Vintern-3B (Doan et al., 2024) further extend these multimodal capabilities to low-resource languages. However, systematic evaluation of these models on Vietnamese chart and table VQA remains limited, primarily due to the lack of suitable high-quality benchmarks in Vietnamese.

Finally, developing VQA datasets for Vietnamese is challenging because of limited multimodal resources and annotations (Nguyen et al., 2023). This scarcity restricts both model training and reliable evaluation. Recent studies show that large language models can assist data annotation and synthesis, reducing human effort while maintaining scalability (Tan et al., 2024; Rocha et al., 2025). In vision–language tasks, LLM-assisted pipelines combined with human verification have proven effective for dataset construction (Choi et al., 2025a). Motivated by these findings, we adopt a semi-automatic, human-in-the-loop annotation strategy for Vietnamese chart and table

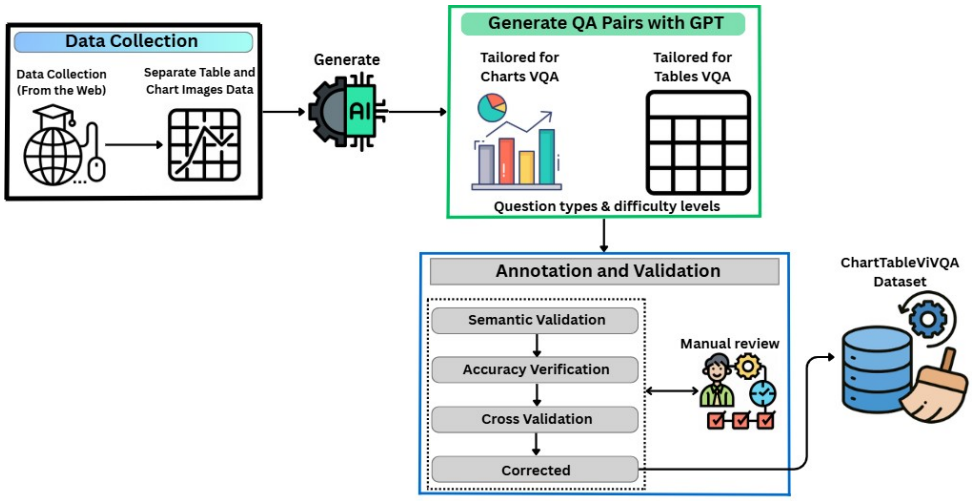


Figure 1: ChartTableViVQA dataset construction pipeline.

VQA. Initial questions, answers, and instructions are generated using LLMs and then reviewed by human annotators to ensure visual grounding and numerical correctness.

3 Dataset

ChartTableViVQA is constructed as a Vietnamese VQA benchmark targeting two types of structured visuals: charts and tables. Each sample consists of an image, a Vietnamese question, a short answer, and a concise instruction describing the reasoning steps required to derive the answer. The dataset aims to cover diverse visual formats, real-world topics, multiple reasoning types (lookup, aggregation, comparison, and trend analysis), and natural Vietnamese language.

3.1 Data Collection

Chart and table images were collected from three major Vietnamese educational platforms—*Vuihoc.vn*, *Vietjack.com*, and *Yourhomework.net* —which frequently reuse charts and tables sourced from official statistical publications, particularly the *Niên giám Thống kê Việt Nam* (Statistical Yearbook of Viet Nam) of the General Statistics Office (Tổng cục Thống kê). As a result, the dataset reflects standardized real-world statistical data across diverse domains such as population, trade, agriculture, and education.

Approximately 920 images were initially collected across subjects such as mathematics, economics, and data analysis. After removing irrelevant, low-quality, and duplicate samples, the final dataset contains 868 images, including 395 charts and 473 tables.

3.2 Iterative Prompt Design

We iteratively refined a Vietnamese meta-prompt to generate question–answer–instruction triplets for each image. Each iteration produced approximately 16 questions covering four reasoning types: lookup, aggregation, comparison, and trend analysis, with constraints progressively added to improve clarity and numerical accuracy. The full prompt templates are provided in Appendix A.

- **Prompt 1 (Initial Version)** The first version defined the basic structure and categories of the 16 questions but lacked diverse phrasing, unit consistency, and clear reasoning steps.

- **Prompt 2 (Refined for Diversity and Formatting)** The second version introduced varied phrasing, standardized year formatting (e.g., "Năm 2019"), and unit consistency. Answers were required to use specific units such as USD or %.

- **Prompt 3 (Added Reasoning and Instruction Format)** The third version required the inclusion of explicit reasoning steps for calculations such as differences and percentages, ensuring consistency between the questions, answers, and data values.

- **Prompt 4 (Final Consolidated Version)** The final version consolidated all improvements, enforcing 100% calculation accuracy, consistent formatting, strict unit handling, and explicit instructions for each question.

Split	Chart		Table	
	Images	Questions	Images	Questions
Training	275	3,902	333	5,219
Develop	60	840	70	1,111
Test	60	826	70	1,105
Total	395	5,568	473	7,435

Table 1: Dataset statistics for each split.

3.3 Question–Answer and Instruction Generation

Using the finalized meta-prompt, we employ ChatGPT (gpt-4o) to generate question–answer–instruction triplets for each chart or table image. For every image, the model produces approximately 16 questions spanning four reasoning types, together with a short answer and a concise instruction.

Questions are designed to assess different reasoning skills over structured visuals, while the accompanying instructions describe the procedural steps required to derive the answer, such as locating relevant values, performing numerical operations, or identifying trends. Instructions complement the questions without repeating their content.

3.4 Annotation and Validation

All automatically generated question–answer–instruction triplets were manually validated by two Vietnamese annotators following a standardized guideline to ensure semantic consistency and numerical correctness.

Annotation Procedure Annotators followed a three-stage validation process:

1. **Semantic consistency check**: verify grammaticality, clarity, and alignment between the question and the visual content. Questions referring to nonexistent entities or mismatched concepts were removed.

2. **Numerical verification**: manually recompute all arithmetic operations (e.g., sums, averages, percentage changes) and correct the answer and instruction when needed.

3. **Cross-review**: overlapping subsets were independently reviewed by multiple annotators, with disagreements resolved through discussion to improve consistency.

Error Categories Annotation errors fall into three main types. **(1) Severe semantic errors** are discarded due to incompatibility with the visual data. **(2) Repairable semantic issues** involve unclear or awkward phrasing and are corrected while preserving the original intent. **(3) Computational errors** arise from incorrect numerical calculations and are fixed through manual recomputation. Representative examples for each category are provided in Appendix B.

Overall, from the 13,904 question triplets automatically generated by LLMs, 901 triplets (6.5%) were completely discarded during validation. The remaining 13,003 triplets (93.5%) passed the automatic validation process. Among these, 5,314 questions (38.2%) required manual editing. Of the edited cases, 2,226 were revised for semantic or phrasing issues, and 3,088 were corrected for numerical accuracy. This indicates that quantitative verification was the dominant component of the annotation effort.

Data split To properly evaluate model performance and prevent visual information leakage, we partition the dataset into training, develop and test sets strictly at the image level. As detailed in Table 1, the partitioning roughly follows a 70-15-15 ratio.

3.5 Dataset Analysis

We categorize questions into three difficulty levels based on the reasoning complexity required to derive the answer:

- **Easy**: Lookup and Image Info, which only require reading a single value or identifying the overall content of a chart or table.

- **Medium**: Aggregation, involving basic arithmetic operations such as sums, differences, averages over multiple values.

- **Hard**: Comparison and Trend Analysis, which require comparing values across categories or time periods (e.g., determining which is higher/lower and by how much) and analyzing increases or decreases over time.

Figure 2 shows the distribution of question types in the ChartVQA and TableVQA subsets of ChartTableViVQA, where minor categories are merged into Aggregation. In both subsets, Lookup accounts for the largest proportion of questions (46.47% for ChartVQA and 55.87% for

TableVQA), followed by Comparison and Trend Analysis, while Aggregation and Image Info make up smaller shares. This reflects our design choice to emphasize multi-value reasoning rather than direct value retrieval. For chart images, hard questions dominate across all categories, peaking in Stacked Area Charts (92.9%) and Area Charts (85.9%). Medium-level questions (Aggregation) are notably more common in chart types with multiple interacting data series, such as Stacked Bar Charts (35.7%), Combo Charts (26.8%), and Grouped Bar Charts (22.7%). Easy questions (Lookup and Image Info) remain the minority across all types (4.8%–8.4%), suggesting that the structural characteristics of each chart type naturally shape the reasoning demands of the associated questions. The details of the difficulty of question distribution by chart types are shown in as shown in Appendix C.

	Question	Answer	Instruction
Max length	35	42	120
Average length	16.08	4.07	25.60
Min length	5	1	4
Median length	16	3	26

Table 2: Token length statistics of questions, answers, and instructions in ChartTableViVQA.

Averaging across ChartVQA and TableVQA yields 15.42% easy, 51.17% medium, and 33.42% hard questions. This stratification is used in Section 5 to analyze model performance by difficulty level. Appendix F presents several representative examples covering all major question types. Additionally, Table 2 reports token-length statistics: questions are moderately long (16.08 tokens on average), answers are short (4.07 tokens), and instructions are longer (25.60 tokens), reflecting the need for explicit step-by-step reasoning over charts and tables.

Finally, each instance in ChartTableViVQA is stored as a unified JSON object containing image-level metadata, a structured representation of the underlying chart or table data, and a list of question–answer–instruction triplets. Full examples of ChartQA and TableQA are illustrated in the Appendix G and H.

4 Methodology

4.1 Task formulation and prompting

Each instance in ChartTableViVQA is represented as a quadruple (I, q, s, a), where I is the chart or table image, q is the Vietnamese question, s is an optional step-by-step instruction, and a is the gold answer. We evaluate all models under two settings: *zero-shot* (input: only I and q) and *instruction-based* (input: I, q, and s). In both settings, models generate a Vietnamese answer $\hat{a}$ using a unified prompt format. Gold answers are intentionally short and unit-sensitive to support precise automatic evaluation of numerical and visual reasoning (Section 4.3).

4.2 Vision–Language Models

We evaluate a comprehensive suite of vision-language models covering both closed-source and open-source architectures, diverse parameter scales, and varying levels of Vietnamese specialization, including:

Closed-source models include the frontier proprietary systems **GPT-4o** (OpenAI et al., 2024) and **Gemini-2.5-Flash** (Comanici et al., 2025).

Open-source models comprise general multilingual systems such as **Qwen2.5-VL-7B-Instruct** (Bai et al., 2023), **Qwen3-VL-8B-Instruct**, **Qwen3-VL-8B-Thinking** (Bai et al., 2025), and **Llama-3.2-11B-Vision-Instruct** (Grattafiori et al., 2024), together with Vietnamese-centric models **LaVy-7B** (Tran and Thanh, 2024) and **Vintern-3B** (Doan et al., 2024). We additionally include the open-weight baseline **Unicorn-VL-R3**[1] for broader coverage.

This selection enables a balanced comparison between proprietary and community-driven models, between general-purpose and language-specific designs, and across different model scales.

4.3 Evaluation metrics

We evaluate model predictions on ChartTableViVQA by comparing the generated answer $\hat{a}$ with the gold answer a using both automatic metrics and a strong LLM-based judge.

BLEU (Papineni et al., 2002): measures n-gram precision between the predicted and reference answers. We report BLEU-1 to BLEU-4 with uniform weights and smoothing, which are suitable for

[1] https://huggingface.co/unicorn-team/Unicorn-VL-R3

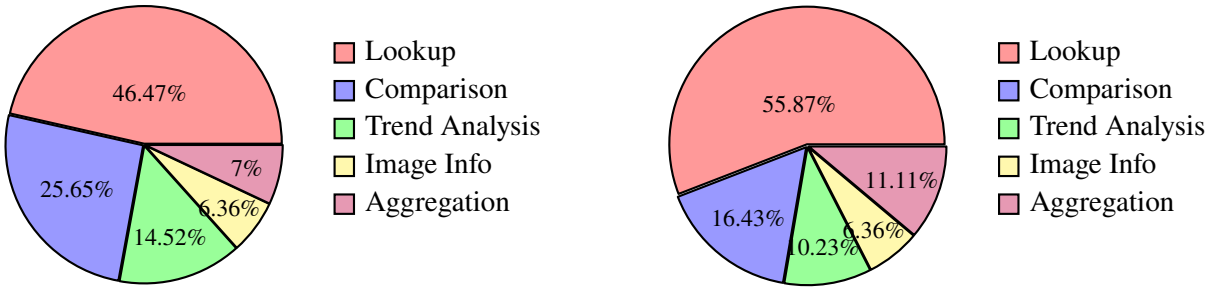


Figure 2: Question type distribution in the ChartVQA (left) and TableVQA (right) datasets.

short, factual responses where exact token matching is important.

ROUGE (Lin, 2004): a recall-oriented overlap metric. We report ROUGE-1 and ROUGE-2 to assess how well the predicted answers cover the content of the gold answers.

METEOR (Banerjee and Lavie, 2005): combines unigram precision and recall. Due to limited Vietnamese support, we compute METEOR using exact word matches only.

BERTScore (Zhang et al., 2019): measures semantic similarity using contextualised multilingual embeddings, enabling soft matching beyond exact token overlap.

G-Eval (Liu et al., 2023): We use GPT-4o as a reference-based LLM judge to evaluate the generated answer $\hat{a}$ by directly comparing it with the ground-truth answer a. The judge scores the prediction on a scale from 0 to 1 based on factual correctness, completeness, and relevance. This metric is reported in percentages.

5 Experimental Results

5.1 Overall Results on Table and Chart Questions

Tables 3 and 4 present the performance of various vision-language models on the table test set (1,105 questions) and the chart test set (826 questions) of the ChartTableViVQA benchmark under two settings: instruction-based and zero-shot (no instruction). The details of implementation configurations are described in Appendix D.

On the **table test set** (Table 3), Qwen models achieve the most outstanding results in the instruction-based setting. **Qwen3-VL-8B-Instruct** leads in BERTScore with 91.2%. Meanwhile, **Qwen2.5-VL-7B-Instruct** demonstrates the best generation quality with BLEU-4 of 39.1%, ROUGE-2 of 79.2%, METEOR of 73.9%, and notably the highest **LLM Judge** score (95.3%). **GPT-4o** ranks closely behind with a BERTScore of 90.3% and LLM Judge of 92.6%. In the zero-shot setting, **GPT-4o** leads in LLM Judge (91.3%), while **Qwen2.5-VL-7B-Instruct** maintains the highest BERTScore (89.5%). Vietnamese-focused models such as **LaVy-7B** and **Vintern-3B** show significantly lower performance.

On the **chart test set** (Table 4), **GPT-4o** establishes clear leadership in both settings, achieving the highest BERTScore (92.3% with instruction and 90.7% zero-shot). **Qwen2.5-VL-7B-Instruct** remains highly competitive, particularly recording the highest LLM Judge score (94.8%) in the instruction-based setting. Other models such as Gemini-2.5-Flash, Llama-11B, LaVy-7B, and Vintern-3B fall into the mid-to-lower tier, while Qwen3-VL-8B-Thinking and Unicorn-VL-R3 remain at the bottom.

The results reveal a significant modality gap between the two data types. All models perform substantially better on structured tabular data than on visually complex charts, with sharp declines in BLEU-4, ROUGE-2, METEOR, and LLM Judge scores. This highlights the greater challenge of chart understanding, which requires advanced visual analysis, axis and legend interpretation, spatial reasoning, and quantitative trend extraction.

The use of instruction-based prompting brings remarkable improvements for the top-performing models. On the table test set, **Qwen2.5-VL-7B-Instruct** improves its LLM Judge from 64.7% to 95.3% (an increase of 30.6%), and METEOR from 59.3% to 73.9% (an increase of 14.6%). On the chart test set, the model significantly boosts its LLM Judge from 57.5% to 94.8% (an increase of 37.3%) and METEOR from 39.4% to 65.9% (an increase of 26.5%). **GPT-4o** also shows notable gains in both datasets. These results confirm the critical importance of detailed instructional guid-

Model	BERTScore	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	METEOR	G-Eval
Instruction-based									
Qwen3-VL-8B-Instruct	**91.2**	66.2	57.7	54.0	37.2	83.6	77.5	69.9	92.6
Qwen2.5-VL-7B-Instruct	90.8	**69.6**	**62.7**	**56.6**	**39.1**	**84.3**	**79.2**	**73.9**	**95.3**
GPT-4o	90.3	66.0	59.0	53.6	34.8	82.2	76.7	69.8	92.6
Llama-11B	81.2	48.8	44.6	40.1	28.1	64.5	56.8	54.3	78.4
Gemini-2.5-Flash	80.3	38.4	33.9	29.4	19.6	62.4	51.3	42.0	84.1
Vintern-3B	79.1	29.1	20.0	16.3	13.1	61.8	48.3	35.6	88.0
LaVy-7B	76.6	27.6	23.9	18.3	14.5	46.5	40.1	58.8	94.6
Qwen3-VL-8B-Thinking	67.7	18.6	15.7	14.1	8.8	31.4	26.1	26.1	79.2
Unicorn-VL-R3	56.6	0.7	0.5	0.3	0.2	7.5	4.5	6.1	85.0
Zero-shot (No Instruction)									
Qwen2.5-VL-7B-Instruct	**89.5**	56.5	48.0	37.6	27.1	69.1	61.0	59.3	64.7
GPT-4o	88.7	**56.7**	**51.5**	**44.1**	**29.7**	**75.0**	**65.8**	**61.5**	**91.3**
Qwen3-VL-8B-Instruct	77.9	20.0	15.3	12.0	9.6	41.8	30.2	24.9	57.2
Gemini-2.5-Flash	73.5	19.5	15.6	11.5	7.8	48.0	33.7	23.2	78.2
Llama-11B	71.6	19.2	13.4	8.7	6.5	32.2	20.3	18.7	28.5
Vintern-3B	71.6	14.2	9.3	6.5	4.8	36.9	25.8	21.2	62.1
LaVy-7B	70.6	12.6	7.4	4.4	3.2	28.3	17.7	29.3	24.1
Qwen3-VL-8B-Thinking	69.4	18.5	15.4	12.2	8.1	35.3	28.3	25.3	77.6
Unicorn-VL-R3	62.4	4.1	2.9	2.2	1.4	11.4	6.7	5.0	46.4

Table 3: Results on the table test set (1,105 questions) with and without instructions (all scores in %).

Model	BERTScore	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	METEOR	G-Eval
Instruction-based									
GPT-4o	**92.3**	**69.3**	**50.2**	**40.5**	**26.6**	**79.8**	**62.8**	**69.5**	90.2
Qwen2.5-VL-7B-Instruct	89.7	65.5	48.1	39.6	27.0	76.8	61.7	65.9	**94.8**
Qwen3-VL-8B-Instruct	88.8	60.8	42.7	36.2	23.9	76.5	58.5	61.2	88.4
Gemini-2.5-Flash	84.1	49.7	30.6	21.9	15.3	69.6	47.9	48.5	83.1
LaVy-7B	83.4	49.5	37.4	24.6	18.0	62.2	48.7	62.6	95.8
Llama-11B	82.4	47.9	35.6	29.8	20.6	62.8	48.0	51.2	82.7
Vintern-3B	79.3	33.8	21.5	16.1	12.4	51.6	39.2	35.5	80.7
Qwen3-VL-8B-Thinking	70.4	25.2	18.0	15.8	10.1	36.5	28.8	29.1	78.1
Unicorn-VL-R3	56.9	1.2	0.8	0.6	0.4	6.0	3.2	6.6	90.5
Zero-shot (No Instruction)									
GPT-4o	**90.7**	**58.7**	**42.2**	**32.7**	**21.8**	**69.1**	**52.1**	**57.8**	**76.3**
Qwen2.5-VL-7B-Instruct	82.6	37.0	27.5	21.1	14.6	48.9	37.0	39.4	57.5
Gemini-2.5-Flash	81.5	39.4	22.2	15.5	11.1	60.1	38.8	38.0	77.7
Qwen3-VL-8B-Instruct	81.4	29.9	17.4	12.8	9.5	45.9	29.2	31.4	60.6
Llama-11B	78.0	24.9	16.9	11.6	8.2	36.3	23.5	24.2	39.9
Vintern-3B	74.8	16.2	9.7	6.6	5.2	26.7	16.0	17.6	44.2
LaVy-7B	74.6	18.3	9.7	6.3	5.0	34.3	21.8	21.9	30.1
Qwen3-VL-8B-Thinking	73.4	20.1	14.4	10.2	7.0	39.1	28.6	22.4	72.5
Unicorn-VL-R3	69.6	14.6	10.4	7.3	5.1	28.9	21.3	16.2	60.5

Table 4: Results on the chart test set (826 questions) with and without instructions (all scores in %).

Subset	Question type	Count	BERTScore (%)	BLEU-1 (%)	BLEU-2 (%)	BLEU-3 (%)	BLEU-4 (%)	ROUGE-1 (%)	ROUGE-2 (%)	METEOR (%)	G-Eval (%)
Chart	Image Info	26	90.8	71.6	**66.6**	**61.5**	**57.1**	84.3	77.8	**84.7**	95.5
Chart	Lookup	196	93.9	79.9	57.3	52.3	31.9	85.7	70.6	78.6	97.9
Chart	Aggregation	138	**95.0**	**82.4**	63.8	56.6	35.4	**89.7**	**79.3**	82.3	**98.7**
Chart	Comparison	258	85.9	53.5	39.9	29.0	21.1	65.9	48.8	55.9	90.1
Chart	Trend Analysis	208	86.9	54.7	36.7	26.7	20.6	72.4	55.9	53.0	95.1
Table	Image Info	70	95.9	85.5	83.2	81.0	**78.4**	93.3	90.3	90.4	96.4
Table	Lookup	189	**98.1**	**92.8**	**87.7**	**81.5**	50.2	**98.2**	**95.3**	**94.5**	98.9
Table	Aggregation	351	96.3	87.2	80.2	76.0	48.9	93.2	90.7	90.5	**99.0**
Table	Comparison	208	75.0	27.1	26.4	15.5	11.1	59.0	50.7	42.5	87.7
Table	Trend Analysis	287	89.5	59.8	46.3	40.3	30.4	80.4	72.4	58.8	93.8

Table 5: Average scores by question type for Qwen2.5-VL-7B-Instruct on the chart (826 questions) and table (1,105 questions) test subsets with instructions.

ance in enhancing the structured visual reasoning capabilities of current vision-language models.

5.2 Results by Question Type and Difficulty

To provide a finer-grained analysis, we break down the performance of **Qwen2.5-VL-7B-Instruct**

according to our question taxonomy in the instruction-based setting (Table 5). Following the defined difficulty hierarchy—*Image Info* and *Lookup* (easy), *Aggregation* (medium), and *Comparison* and *Trend Analysis* (hard)—the results strongly validate the intended design of the benchmark. Easy questions are largely solved across both modalities. On tables, *Image Info* achieves a BLEU-4 of 78.4%, METEOR of 90.4%, and LLM Judge of 96.4%, while *Lookup* attains 50.2% BLEU-4, 94.5% METEOR, and 98.9% LLM Judge. On charts, *Image Info* reaches 57.1% BLEU-4, 84.7% METEOR, and 95.5% LLM Judge, and *Lookup* obtains 31.9% BLEU-4, 78.6% METEOR, and 97.9% LLM Judge. These strong scores reflect robust visual grounding and precise value extraction capabilities. Medium-difficulty *Aggregation* questions also yield excellent results, particularly on tables (BLEU-4 48.9%, METEOR 90.5%, LLM Judge 99.0%) and to a lesser extent on charts (BLEU-4 35.4%, METEOR 82.3%, LLM Judge 98.7%). In contrast, hard questions requiring complex numerical reasoning remain significantly more challenging. On tables, *Comparison* is the most difficult category (BLEU-4 11.1%, METEOR 42.5%, LLM Judge 87.7%), followed by *Trend Analysis* (BLEU-4 30.4%, METEOR 58.8%, LLM Judge 93.8%). On charts, both hard categories show notable weakness: *Comparison* at 21.1% BLEU-4, 55.9% METEOR, 90.1% LLM Judge and *Trend Analysis* at 20.6% BLEU-4, 53.0% METEOR, 95.1% LLM Judge. Overall, the clear performance stratification across difficulty levels confirms the taxonomy's validity. While **Qwen2.5-VL-7B-Instruct** excels at straightforward visual-text alignment and simple retrieval tasks (reflected in very high LLM Judge scores), tasks involving comparative reasoning, multi-step aggregation, and temporal/spatial trend interpretation continue to expose the current limitations of vision-language models, especially in the chart modality.

5.3 Error Analysis

Subset	#Q	Numerical	Mismatch	Comparison
Chart	151	5	41	11
Table	159	4	55	18

Table 6: Distribution of error types for Qwen2.5-VL-7B-Instruct on a sampled set of chart and table questions in the instruction setting.

We analyze the failure modes of Qwen2.5-VL-7B-Instruct by manually inspecting its predictions on 151 chart questions and 159 table questions in the instruction-based setting. Errors are classified into three main categories: Mismatch, Comparison, and Numerical. The distribution of these error types is summarized in Table 6, with representative examples provided in the Appendix E. Mismatch errors are the most prevalent in both modalities, accounting for 41 out of 57 errors on charts and 55 out of 77 errors on tables. These cases typically involve producing a semantically correct numerical value but failing to match the expected surface form (e.g., differences in sign notation, decimal formatting, or unit presentation). This highlights how strict string-based evaluation metrics can penalize predictions that are conceptually accurate yet unnormalized. Comparison errors form the second largest category, with 11 instances on charts and 18 on tables. Here, the model correctly extracts relevant values but fails to infer the proper relational conclusion (e.g., misjudging which value is higher, lower, or equal). This pattern is consistent with the notably low BLEU-4 and METEOR scores observed for Comparison questions in Table 5, underscoring ongoing difficulties in relational reasoning even under guided prompting. Pure numerical errors are the least common (5 on charts and 4 on tables) and primarily stem from mistakes in magnitude, scaling, or unit interpretation (e.g., treating scaled values as absolute quantities). Overall, the qualitative error analysis complements the automatic metrics by revealing that answer normalization issues and comparison-based reasoning remain the dominant sources of failure for Qwen2.5-VL-7B-Instruct on ChartTableViVQA, despite step-by-step instructions usage.

6 Conclusion

This paper introduced **ChartTableViVQA**, a benchmark dataset for Vietnamese Visual Question Answering over chart and table images. The dataset consists of 13K human-validated question-answer-instruction triplets that span multiple reasoning types and difficulty levels, providing a systematic evaluation setting for a low-resource language. Through extensive benchmarking, we showed that current vision-language models perform substantially better on tables than on charts, and that instruction-based prompting consistently improves performance. Finally, ChartTableViVQA facilitates future research on model adaptation,

reasoning-aware architectures, and training strategies for chart and table understanding in low-resource settings. The dataset and source code are available at https://github.com/sonlam1102/charttablevivqa.

Limitations

Despite the improvements brought by instruction-based prompting, current vision–language models still exhibit fundamental limitations in numerical reasoning on ChartTableViVQA. Although models can often identify relevant numerical values from tables or charts, their reasoning remains unstable across similar questions, resulting in inconsistent predictions for numerically comparable cases.

Relational operators such as higher, lower, increase, decrease, and difference are not consistently grounded in correct numerical comparisons, especially when multiple reference values or directional changes must be considered. These issues are particularly evident in comparison and trend analysis tasks, where correct answers require reasoning over relationships rather than direct value extraction.

Furthermore, model performance degrades on multi-step and multi-reference questions, such as aggregations across rows or columns or analyses over multiple time points. Even when intermediate values are correctly identified, failures in combining them coherently highlight the lack of reliable structured and compositional numerical reasoning in current vision–language models.

Acknowledgments

This research was supported by The VNUHCM-University of Information Technology's Scientific Research Support Fund.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2020. 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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juhwan Choi, Junehyoung Kwon, JungMin Yun, Seunguk Yu, and YoungBin Kim. 2025a. Voldoger: Llm-assisted datasets for domain generalization in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6833–6843.

Juhwan Choi, Junehyoung Kwon, JungMin Yun, Seunguk Yu, and YoungBin Kim. 2025b. Voldoger: Llm-assisted datasets for domain generalization in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6833–6843.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the*

North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pages 4171–4186.

Khang T Doan, Bao G Huynh, Dung T Hoang, Thuc D Pham, Nhat H Pham, Quan Nguyen, Bang Q Vo, and Suong N Hoang. 2024. Vintern-1b: an efficient multimodal large language model for vietnamese. *arXiv preprint arXiv:2408.12480*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shamanthak Hegde, Pooyan Fazli, and Hasti Seifi. 2025. Chartqa-x: Generating explanations for charts. *arXiv preprint arXiv:2504.13275*.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1918–1925, Marseille, France. European Language Resources Association.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

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

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Carlos Rocha, Jonatas S Grosman, Fernando A Correia, Venicius Rego, and Hélio Lopes. 2025. A data annotation approach using large language models. In *ICEIS (1)*, pages 748–755.

Hui Shi, Yusheng Xie, Luis Goncalves, Sicun Gao, and Jishen Zhao. 2024. Wikidt: Visual-based table recognition and question answering dataset. In *International Conference on Document Analysis and Recognition*, pages 406–437. Springer.

Brandon Smock, Rohith Pesala, and Robin Abraham. 2022. 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.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Chi Tran and Huong Le Thanh. 2024. Lavy: Vietnamese multimodal large language model. *arXiv preprint arXiv:2404.07922*.

Tổng cục Thống kê. Niên giám thống kê việt nam. https://www.nso.gov.vn/nien-giam/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9102–9124, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix: Full Prompt Templates with Example Outputs

Below are the full versions of each Vietnamese meta-prompt used in the experiments. Each version is followed by example outputs and their English translations.

A.1 Prompt 1 – Initial Version (Basic Questions)

Prompt in Vietnamese (Prompt 1): Dựa trên ảnh Bảng/Biểu đồ trên, hãy tạo 16 câu hỏi cho một bảng/biểu đồ số liệu (bạn tự chọn), chia thành 4 nhóm, mỗi nhóm 4 câu. Các nhóm câu hỏi bao gồm:

1. **Truy xuất dữ liệu trực tiếp (lookup)** – 4 câu Mỗi câu yêu cầu xác định rõ hàng và cột chứa dữ liệu và chỉ truy xuất một giá trị cụ thể từ bảng. Không yêu cầu tính toán.
2. **Phép toán cơ bản (aggregation)** – 4 câu Mỗi câu yêu cầu xác định phạm vi dữ liệu và nêu rõ phép toán cần áp dụng như tổng (sum), trung bình (average), hoặc hiệu (difference). Kết quả phải được tính toán chính xác.
3. **So sánh (comparison)** – 4 câu Mỗi câu yêu cầu so sánh hai hoặc nhiều giá trị từ bảng, với yêu cầu rõ ràng về loại so sánh cần thực hiện (lớn nhất, nhỏ nhất, cao hơn, thấp hơn). Kết quả phải có đơn vị cụ thể và chỉ rõ sự chênh lệch.
4. **Phân tích xu hướng (trend analysis)** – 4 câu Câu hỏi yêu cầu phân tích sự thay đổi của giá trị theo thời gian. Ví dụ: "Từ năm 2015 đến năm 2021, GDP của Ma-lai-xi-a tăng hay giảm bao nhiêu %?"

Yêu cầu:

- Có một câu hỏi tổng quan về nội dung bảng, như "Biểu đồ trên cho thấy gì?"
- Câu hỏi cần rõ ràng và dễ hiểu

English translation (Prompt 1): Based on the table/chart image provided, create 16 questions for any data table or chart you choose, divided into 4 groups of 4 questions each. The groups are:

1. **Data Retrieval (lookup)** – 4 questions Each question requires identifying the row and column containing the data and retrieving a specific value from the table, with no calculations.
2. **Basic Operations (aggregation)** – 4 questions Each question requires specifying a data range and the operation to perform, such as sum, average, or difference. Results must be calculated accurately.
3. **Comparison** – 4 questions Each question requires comparing two or more values from the table, with clear instructions on the type of comparison (largest, smallest, higher, lower). Results must include specific units and the difference.
4. **Trend Analysis** – 4 questions Each question requires analyzing the change in values over time. Example: "From 2015 to 2021, did Malaysia's GDP increase or decrease by what %?"

Requirements:

- At least one question should summarize the overall content of the table/chart.
- Questions should be clear and easy to understand.

Example Output for Prompt 1 (Initial Version)

1. **Question:** Biểu đồ nói về gì?
   Answer: Thu nhập bình quân đầu người của Việt Nam và Singapore từ 2015 đến 2020.
2. **Question:** Năm 2020 Việt Nam bao nhiêu?
   Answer: 2779
3. **Question:** Singapore năm 2018 là bao nhiêu?
   Answer: 66679
4. **Question:** Tổng thu nhập Việt Nam 6 năm là bao nhiêu?
   Answer: khoảng 15000

English Translation (Example Output for Prompt 1):

1. **Question:** What does the chart show?
   Answer: The per capita income of Vietnam and Singapore from 2015 to 2020.
2. **Question:** How much was Vietnam's income in 2020?
   Answer: 2779
3. **Question:** How much was Singapore's income in 2018?
   Answer: 66679

4. **Question:** What is the total income of Vietnam over 6 years?
   Answer: approximately 15000

A.2 Prompt 2 – Refined Version

Prompt in Vietnamese (Prompt 2): Dựa trên ảnh Bảng/ Biểu đồ trên, hãy tạo 16 câu hỏi cho một bảng/biểu đồ số liệu (bạn tự chọn), chia thành 4 nhóm, mỗi nhóm 4 câu. Các nhóm câu hỏi bao gồm:

1. **Truy xuất dữ liệu trực tiếp (lookup)** – 4 câu Mỗi câu yêu cầu xác định rõ hàng và cột chứa dữ liệu và chỉ truy xuất một giá trị cụ thể từ bảng. Không yêu cầu tính toán.

2. **Phép toán cơ bản (aggregation)** – 4 câu Mỗi câu yêu cầu xác định phạm vi dữ liệu và nêu rõ phép toán cần áp dụng như tổng (sum), trung bình (average), hoặc hiệu (difference). Kết quả phải được tính toán chính xác.

3. **So sánh (comparison)** – 4 câu Mỗi câu yêu cầu so sánh hai hoặc nhiều giá trị từ bảng, với yêu cầu rõ ràng về loại so sánh cần thực hiện (lớn nhất, nhỏ nhất, cao hơn, thấp hơn). Kết quả phải có đơn vị cụ thể và chỉ rõ sự chênh lệch.

4. **Phân tích xu hướng (trend analysis)** – 4 câu Câu hỏi yêu cầu phân tích sự thay đổi của giá trị theo thời gian. Ví dụ: "Từ năm 2015 đến năm 2021, GDP của Ma-lai-xi-a tăng hay giảm bao nhiêu %?"

Yêu cầu:

- Có một câu hỏi tổng quan về nội dung bảng, như "Biểu đồ trên cho thấy gì?".
- Câu hỏi đa dạng cách hỏi, luôn thay đổi chủ ngữ, tránh lặp lại.
- Với năm, trả lời thêm "Năm" trước, như "Năm 2019" thay vì chỉ 2019.
- Thêm đa dạng cụm từ như "tăng/giảm bao nhiêu phần trăm", "cao hơn hay thấp hơn".
- Đáp án phải có đơn vị đầy đủ.

English translation (Prompt 2): Based on the table/chart image provided, create 16 questions for any data table or chart you choose, divided into 4 groups of 4 questions each. The groups are:

1. **Data Retrieval (lookup)** – 4 questions Each question requires identifying the row and column containing the data and retrieving a specific value from the table, with no calculations.

2. **Basic Operations (aggregation)** – 4 questions Each question requires specifying a data range and the operation to perform, such as sum, average, or difference. Results must be calculated accurately.

3. **Comparison** – 4 questions Each question requires comparing two or more values from the table, with clear instructions on what type of comparison is needed (largest, smallest, higher, lower). Results must include specific units and the difference.

4. **Trend Analysis** – 4 questions Each question requires analyzing the change in values over time. Example: "From 2015 to 2021, did Malaysia's GDP increase or decrease by what %?"

Requirements:

- At least one question should summarize the overall content of the table/chart.
- Questions should vary in phrasing, always changing the subject and avoiding repetition.
- For year-related questions, include "Year" before the year, such as "Year 2019" instead of just 2019.
- Add variety in phrases such as "increase/decrease by how much", "higher or lower", "how much did it change", etc.
- Answers must always include full units (e.g., USD, %, million VND, etc.).

Example Output for Prompt 2 (Refined Version)

1. **Question:** Biểu đồ trên thể hiện nội dung gì?
   Answer: Thu nhập bình quân đầu người/năm của Việt Nam và Singapore từ 2015 đến 2020.

2. **Question:** Năm 2020, thu nhập bình quân đầu người của Việt Nam là bao nhiêu?
   Answer: 2779 đô la Mỹ.

3. **Question:** Thu nhập bình quân đầu người của Singapore năm 2020 tăng hay giảm bao nhiêu phần trăm so với năm 2019?
   Answer: Giảm khoảng 8,3%.

4. **Question:** Năm 2019, thu nhập bình quân đầu người của Việt Nam cao hơn hay thấp hơn năm 2018 bao nhiêu đô la Mỹ?
   Answer: Cao hơn 144 đô la Mỹ.

English Translation (Example Output for Prompt 2):

1. **Question:** What does the chart show?
   Answer: The per capita income per year of Vietnam and Singapore from 2015 to 2020.

2. **Question:** What was the per capita income of Vietnam in 2020?
   Answer: 2779 USD.

3. **Question:** Did Singapore's per capita income in 2020 increase or decrease by how much compared to 2019?
   Answer: Decreased by approximately 8.3%.

4. **Question:** Was Vietnam's per capita income in 2019 higher or lower than in 2018 by how much?
   Answer: Higher by 144 USD.

A.3 Prompt 3 – Enhanced Version with Instruction

Prompt in Vietnamese (Prompt 3): Dựa trên ảnh Bảng/ Biểu đồ trên, hãy tạo 16 câu hỏi cho một bảng/biểu đồ, chia thành 4 nhóm chính, mỗi nhóm 4 câu. Các nhóm câu hỏi bao gồm:

1. **Truy xuất dữ liệu trực tiếp (lookup)** – 4 câu Mỗi câu yêu cầu xác định rõ hàng và cột chứa dữ liệu và chỉ truy xuất một giá trị cụ thể từ bảng. Không yêu cầu tính toán.

2. **Phép toán cơ bản (aggregation)** – 4 câu Mỗi câu yêu cầu xác định phạm vi dữ liệu và nêu rõ phép toán cần áp dụng như tổng (sum), trung bình (average), hoặc hiệu (difference). Kết quả phải được tính toán chính xác.

3. **So sánh (comparison)** – 4 câu Mỗi câu yêu cầu so sánh hai hoặc nhiều giá trị từ bảng, với yêu cầu rõ ràng về loại so sánh cần thực hiện (lớn nhất, nhỏ nhất, cao hơn, thấp hơn). Kết quả phải có đơn vị cụ thể và chỉ rõ sự chênh lệch.

4. **Phân tích xu hướng (trend analysis)** – 4 câu Câu hỏi yêu cầu phân tích sự thay đổi của giá trị theo thời gian. Ví dụ: "Từ năm 2015 đến năm 2021, GDP của Ma-lai-xi-a tăng hay giảm bao nhiêu %?"

Yêu cầu:

- Mỗi câu hỏi và đáp án phải có phần **Instruction** chỉ rõ vị trí hàng/cột và cách tính toán (nếu có).
- Năm phải ghi đầy đủ theo định dạng "Năm 20XX".
- Câu hỏi và đáp án phải rõ ràng và chính xác.

Hướng dẫn trả lời và đáp án:

- Đáp án phải có đơn vị rõ ràng (%, tỷ đồng, triệu người, nghìn tấn, km², ... tùy vào từng trường hợp).
- Với "chênh lệch", chỉ ghi giá trị chênh lệch, không cần ghi "cao hơn" hay "thấp hơn".
- Với "cao hơn" hoặc "thấp hơn", chỉ cần ghi mức chênh lệch, không cần thêm từ "chênh lệch".
- Với câu hỏi "thay đổi như thế nào", trả lời theo cấu trúc "tăng" hoặc "giảm" bao nhiêu và kèm đơn vị, ví dụ: "Năm 2019, tỷ lệ thất nghiệp ở thành thị có tăng hay giảm so với năm 2015?".
- Khi câu hỏi yêu cầu tính phần trăm, phải thực hiện phép toán và trả lời theo tỷ lệ phần trăm.

Nguyên tắc Instruction:

- Mỗi câu hỏi phải có hướng dẫn rõ ràng, chỉ rõ vị trí hàng và cột trong bảng.
- Câu truy xuất dữ liệu (lookup) bắt đầu bằng: "Tìm giá trị tại cột 'A', hàng 'B' là ...".
- Câu về chênh lệch (difference) bắt đầu bằng: "Lấy giá trị tại cột 'A', hàng 'B' là số liệu" và tiếp theo là phép trừ để tính hiệu. Ví dụ: "Lấy giá trị tại cột 'Bắc Trung Bộ và Duyên

hải Nam Trung Bộ' tại cột '2021' là 1198.7 và trừ giá trị tại cột '2020' là 1157.7: 1198.7 - 1157.7 = 17.0".

- Câu về so sánh bắt đầu bằng: "So sánh giá trị tại hàng 'B', cột 'A' và hàng 'C', cột 'B'". Sau đó sẽ nêu rõ kết quả và sự chênh lệch.
- Câu hỏi về tăng/giảm thay đổi sử dụng cấu trúc tương tự như câu so sánh nhưng ghi "tăng" hoặc "giảm".
- Câu tính toán (tổng, trung bình, phần trăm,...) cần chỉ rõ cột và hàng, ví dụ: "Tính tổng tỷ lệ thất nghiệp ở thành thị trong các năm từ 2010 đến 2021".

Đồng bộ giữa Instruction và Answer:

- Đảm bảo đáp án phải khớp hoàn toàn với Instruction về số liệu và đơn vị.
- Kiểm tra kỹ lưỡng các phép toán trong Instruction để đảm bảo chúng đúng với dữ liệu bảng.

English translation (Prompt 3): Based on the table/chart image provided, create 16 questions for a table/chart, divided into 4 main groups, each group containing 4 questions. The groups are:

1. **Data Retrieval (lookup)** – 4 questions Each question requires identifying the row and column containing the data and retrieving a specific value from the table, with no calculations.
2. **Basic Operations (aggregation)** – 4 questions Each question requires specifying a data range and the operation to perform, such as sum, average, or difference. Results must be calculated accurately.
3. **Comparison** – 4 questions Each question requires comparing two or more values from the table, with clear instructions on what type of comparison is needed (largest, smallest, higher, lower). Results must include specific units and the difference.
4. **Trend Analysis** – 4 questions Each question requires analyzing the change in values over time. Example: "From 2015 to 2021, did Malaysia's GDP increase or decrease by what %?"

Creation Guidelines:

- Every question and answer must include an **Instruction** specifying the row/column positions and calculation method (if applicable).
- Years must be written in full as "Year 20XX".
- Questions and answers must be clear and accurate.

Answer and Instruction Guidelines:

- Answers must include specific units (%, million tons, million people, etc.).
- For "difference", only the value of the difference should be written, without adding "higher" or "lower".
- For "higher" or "lower", only the difference should be provided, without adding the word "difference".
- For "how much change", answer with "increase" or "decrease" and include the unit, e.g., "In 2019, did the urban unemployment rate increase or decrease compared to 2015?".
- When a question asks for a percentage, the calculation must be done and answered in percentage form.

Instruction Guidelines:

- Every question must have clear instructions specifying row/column positions in the table.
- For data retrieval (lookup), instructions should begin with: "Find the value in column 'A', row 'B' as ...".
- For difference, instructions should begin with: "Take the value in column 'A', row 'B' as data" and perform subtraction. For example: "Take the value for 'Central North and South Central Coast' in column '2021' as 1198.7 and subtract the value in column '2020' as 1157.7: 1198.7 - 1157.7 = 17.0".
- For comparison, instructions should begin with: "Compare the value at row 'B', column 'A' and row 'C', column 'B'". Then, state the result and difference.
- For increase/decrease, use the same structure as comparison but specify "increase" or "decrease".

- For calculations (sum, average, percentage, etc.), clearly specify the rows and columns used, e.g., "Calculate the total unemployment rate in urban areas from 2010 to 2021".

Ensure consistency between Instruction and Answer:

- Ensure the answer matches the instruction completely, including values and units.
- Double-check the calculations in the instruction to ensure they match the table data.

Example Output for Prompt 3 (Appearance of Instruction)

1. **Question:** Biểu đồ trên thể hiện nội dung gì?
   Answer: Biểu đồ trên thể hiện Thu nhập bình quân đầu người/năm của Việt Nam và Singapore từ 2015 đến 2020.
   Instruction: Đọc tiêu đề và chú thích của biểu đồ.

2. **Question:** Năm 2020, thu nhập bình quân đầu người của Việt Nam là bao nhiêu?
   Answer: 2779 đô la Mỹ.
   Instruction: Tìm giá trị tại cột 'Việt Nam' tại Năm '2020' là 2779.

3. **Question:** Thu nhập bình quân của Singapore năm 2020 tăng hay giảm bao nhiêu phần trăm so với năm 2019?
   Answer: Giảm 8,33%.
   Instruction: Tính phần trăm thay đổi giữa năm 2020 và 2019 cho Singapore.

4. **Question:** Năm 2019, thu nhập bình quân đầu người của Việt Nam cao hơn hay thấp hơn năm 2018 bao nhiêu đô la Mỹ?
   Answer: Cao hơn 144 đô la Mỹ.
   Instruction: So sánh giá trị tại cột 'Việt Nam' tại Năm '2019' là 2714 và Năm '2018' là 2570: 2714 - 2570 = 144.

English Translation (Example Output for Prompt 3):

1. **Question:** What does the chart show?
   Answer: The per capita income per year of Vietnam and Singapore from 2015 to 2020.
   Instruction: Read the title and caption of the chart.

2. **Question:** What was the per capita income of Vietnam in 2020?
   Answer: 2779 USD.
   Instruction: Find the value in the column 'Vietnam' for Year '2020' as 2779.

3. **Question:** Did Singapore's per capita income in 2020 increase or decrease by how much compared to 2019?
   Answer: Decreased by 8.33%.
   Instruction: Calculate the percentage change between 2020 and 2019 for Singapore.

4. **Question:** Was Vietnam's per capita income in 2019 higher or lower than in 2018 by how much?
   Answer: Higher by 144 USD.
   Instruction: Compare the value in the column 'Vietnam' for Year '2019' as 2714 and Year '2018' as 2570: 2714 - 2570 = 144 (Higher).

A.4 Prompt 4 – Final Consolidated Version (used in experiments)

Prompt in Vietnamese (Prompt 4): Dựa trên ảnh Bảng/ Biểu đồ trên, hãy tạo 16 câu hỏi cho bảng/biểu đồ, chia thành 4 nhóm chính, mỗi nhóm 4 câu. Các nhóm câu hỏi cụ thể bao gồm:

1. **Truy xuất dữ liệu trực tiếp (lookup)** – 4 câu Mỗi câu yêu cầu xác định rõ hàng và cột chứa dữ liệu và chỉ truy xuất một giá trị cụ thể từ bảng, không yêu cầu tính toán.

2. **Phép toán cơ bản (aggregation)** – 4 câu Mỗi câu yêu cầu xác định phạm vi dữ liệu và nêu rõ phép toán cần áp dụng như tổng (sum), trung bình (average), hoặc hiệu (difference). Kết quả phải được tính toán chính xác.

3. **So sánh (comparison)** – 4 câu Mỗi câu yêu cầu so sánh hai hoặc nhiều giá trị từ bảng, với yêu cầu rõ ràng về loại so sánh cần thực hiện (lớn nhất, nhỏ nhất, cao hơn, thấp hơn). Kết quả phải có đơn vị cụ thể và chỉ rõ sự chênh lệch.

4. **Phân tích xu hướng (trend analysis)** – 4 câu Câu hỏi yêu cầu phân tích sự thay đổi của giá trị theo thời gian. Ví dụ: "Từ năm 2015 đến năm 2021, GDP của Ma-lai-xi-a tăng hay giảm bao nhiêu %?"

Yêu cầu:

- Với các câu hỏi liên quan đến năm, trả lời phải bao gồm đầy đủ "Năm" trước giá trị năm, ví dụ: "Năm 2019" thay vì chỉ ghi "2019".

- Đảm bảo sự đa dạng trong các cụm từ như "tăng/giảm bao nhiêu phần trăm", "cao hơn hay thấp hơn", "thay đổi như thế nào".

Cách trả lời và yêu cầu về phần đáp án:

- Đáp án phải có đơn vị rõ ràng (%, tỷ đồng, triệu người, nghìn tấn, km², ... tùy vào từng trường hợp).
- Đối với các câu có từ "chênh lệch", trả lời chỉ cần giá trị chênh lệch mà không cần thêm "cao hơn" hoặc "thấp hơn".
- Câu hỏi có từ "cao hơn" hoặc "thấp hơn" chỉ cần xác định sự khác biệt giữa hai giá trị mà không cần tính thêm "chênh lệch".
- Các câu hỏi có từ "thay đổi như thế nào" phải trả lời theo hướng "tăng hay giảm bao nhiêu" và kèm đơn vị, ví dụ: "Năm 2019, tỷ lệ thất nghiệp ở thành thị có tăng hay giảm so với năm 2015?".
- Khi câu hỏi yêu cầu tính phần trăm, phải thực hiện phép toán và trả lời theo tỷ lệ phần trăm.

Quy tắc cho phần Instruction:

- Mọi câu hỏi phải có instruction rõ ràng, chỉ rõ vị trí hàng và cột trong bảng.
- Câu truy xuất trực tiếp (lookup) bắt đầu bằng: "Tìm giá trị tại cột 'A', hàng 'B' là ...".
- Câu hỏi về chênh lệch (difference) sẽ bắt đầu bằng: "Lấy giá trị tại cột 'A', hàng 'B' là số liệu" và tiếp theo là phép trừ để tính hiệu. Ví dụ: "Lấy giá trị hàng 'Bắc Trung Bộ và Duyên hải Nam Trung Bộ' tại cột '2021' là 1198.7 và trừ giá trị tại cột '2020' là 1157.7: 1198.7 - 1157.7 = 17.0".
- Câu hỏi về so sánh bắt đầu bằng: "So sánh giá trị tại hàng 'B', cột 'A' và hàng 'C', cột 'B'". Sau đó sẽ nêu rõ kết quả và sự chênh lệch.
- Câu hỏi về tăng/giảm thay đổi sử dụng cấu trúc tương tự như câu so sánh nhưng kết quả phải ghi là "tăng" hoặc "giảm".
- Câu hỏi tính toán (tổng, trung bình, phần trăm,...) cần chỉ rõ cột và hàng, ví dụ: "Tính tổng tỷ lệ thất nghiệp ở thành thị trong các năm từ 2010 đến 2021".

Đảm bảo tính nhất quán giữa Instruction và Answer:

- Đảm bảo answer phải trùng khớp hoàn toàn với instruction, cả giá trị số và đơn vị.
- Kiểm tra kỹ lưỡng trước khi đưa ra đáp án để đảm bảo mọi phép toán trong instruction đều được áp dụng đúng theo bảng dữ liệu.

English translation (Prompt 4): Create exactly 16 questions for the table/chart, divided into 4 main groups, each group containing 4 questions. Each question in the groups will follow specific guidelines to ensure clarity and consistency. The groups are:

1. **Data Retrieval (lookup)** – 4 questions Each question requires clearly identifying the row and column containing the data and retrieving a specific value from the table, without calculations.
2. **Basic Operations (aggregation)** – 4 questions Each question requires identifying a data range and specifying a calculation such as sum (sum), average (average), or difference (difference). Results must be calculated accurately.
3. **Comparison** – 4 questions Each question requires comparing two or more values from the table, specifying the type of comparison needed, such as "largest", "smallest", "higher", "lower", "greater", "smaller". The result of the comparison should include specific units and the difference, if applicable.
4. **Trend Analysis** – 4 questions Each question requires analyzing the change in values over time. Example: "From 2015 to 2021, did Malaysia's GDP increase or decrease by what %?"

Creation Guidelines:

- For year-related questions, always include "Year" before the year, e.g., "Year 2019" instead of just "2019".
- Ensure diversity in phrasing such as "increase/decrease by how much", "higher or lower", "how much did it change".

Answer and Instruction Guidelines:

- Answers must include specific units (%, million tons, million people, etc.).

- For "difference", only the value of the difference should be written, without adding "higher" or "lower".
- For "higher" or "lower", only the difference should be provided, without adding the word "difference".
- For "how much change", answer with "increase" or "decrease" and include the unit, e.g., "In 2019, did the urban unemployment rate increase or decrease compared to 2015?".
- When a question asks for a percentage, the calculation must be done and answered in percentage form.

Instruction Guidelines:

- Every question must have clear instructions specifying row/column positions in the table.
- For data retrieval (lookup), instructions should begin with: "Find the value in column 'A', row 'B' as . . . ".
- For difference, instructions should begin with: "Take the value in column 'A', row 'B' as data" and perform subtraction. For example: "Take the value for 'Central North and South Central Coast' in column '2021' as 1198.7 and subtract the value in column '2020' as 1157.7: 1198.7 - 1157.7 = 17.0".
- For comparison, instructions should begin with: "Compare the value at row 'B', column 'A' and row 'C', column 'B'". Then, state the result and difference.
- For increase/decrease, use the same structure as comparison but specify "increase" or "decrease".
- For calculations (sum, average, percentage, etc.), clearly specify the rows and columns used, e.g., "Calculate the total unemployment rate in urban areas from 2010 to 2021".

Ensure consistency between Instruction and Answer:

- Ensure the answer matches the instruction completely, including values and units.
- Double-check the calculations in the instruction to ensure they match the table data.

Example Output for Prompt 4 (Fully Developed Version with Strict Instructions)

1. **Question:** Biểu đồ trên thể hiện nội dung gì?
   Answer: Biểu đồ trên thể hiện Thu nhập bình quân đầu người/năm của Việt Nam và Singapore từ 2015 đến 2020.
   Instruction: Thu nhập bình quân đầu người/năm của Việt Nam và Singapore từ 2015 đến 2020.
2. **Question:** Năm 2020, thu nhập bình quân đầu người của Việt Nam là bao nhiêu?
   Answer: 2779 đô la Mỹ.
   Instruction: Tìm giá trị tại cột 'Việt Nam' tại Năm '2020' là 2779.
3. **Question:** Thu nhập bình quân của Singapore năm 2020 tăng hay giảm bao nhiêu phần trăm so với năm 2019?
   Answer: Giảm 8,33%.
   Instruction: Tính phần trăm thay đổi bằng cách lấy ((59798 - 65233) / 65233) * 100 = -8,33% (Giảm).
4. **Question:** Năm 2019, thu nhập bình quân đầu người của Việt Nam cao hơn hay thấp hơn năm 2018 bao nhiêu đô la Mỹ?
   Answer: Cao hơn 144 đô la Mỹ.
   Instruction: So sánh giá trị tại cột 'Việt Nam' tại Năm '2019' là 2714 và Năm '2018' là 2570: 2714 > 2570, 2714 - 2570 = 144 (Cao hơn).

English Translation (Example Output for Prompt 4):

1. **Question:** What does the chart show?
   Answer: The per capita income per year of Vietnam and Singapore from 2015 to 2020.
   Instruction: Income per year of Vietnam and Singapore from 2015 to 2020.
2. **Question:** What was the per capita income of Vietnam in 2020?
   Answer: 2779 USD.
   Instruction: Find the value in the column 'Vietnam' for Year '2020' as 2779.
3. **Question:** Did Singapore's per capita income in 2020 increase or decrease by how much compared to 2019?
   Answer: Decreased by 8.33%.
   Instruction: Calculate the percentage change by taking ((59798 - 65233) / 65233) * 100 = -8.33% (Decrease).

4. **Question:** Was Vietnam's per capita income in 2019 higher or lower than in 2018 by how much?
   Answer: Higher by 144 USD.
   Instruction: Compare the value in the column 'Vietnam' for Year '2019' as 2714 and Year '2018' as 2570: 2714 > 2570, 2714 - 2570 = 144 (Higher).

B Appendix: Additional Examples of Annotation Errors

This appendix provides illustrative examples for the three error categories described in Section 3.4. The original question–answer pairs are written in Vietnamese, as in the dataset, and are accompanied by English translations for clarity.

B.1 Severe Semantic Errors (Removed)

Example 1: Conceptually invalid terminology.

- **Underlying chart:** Annual crude oil production (million tonnes) of a country.
- **Original Vietnamese question (generated by ChatGPT):**
  "Sản lượng nuôi trồng dầu thô năm 2014 là bao nhiêu?"
  English translation:
  "What is the production of cultivated crude oil in 2014?"
- **Issue:** The phrase *"nuôi trồng dầu thô"* ("cultivated crude oil") is semantically incorrect: crude oil is extracted, not cultivated like crops or aquaculture products.
- **Resolution:** The question is treated as a severe semantic error and removed from the dataset, since fixing it would effectively create a new question with different meaning.

B.2 Repairable Semantic Issues (Corrected and Retained)

Example 2: Unnatural phrasing in a relative comparison.

- **Underlying table:** Population of Viet Nam and Thailand (millions) in 2021.
- **Original Vietnamese question (generated by ChatGPT):**
  "Dân số của Việt Nam năm 2021 có cao hơn 1,5 lần so với Thái Lan không?"
  English translation:
  "Is the population of Viet Nam in 2021 more than 1.5 times higher than that of Thailand?"
- **Issue:** The comparison is meaningful, but the expression *"cao hơn 1,5 lần"* is stylistically stiff in Vietnamese.
- **Revised Vietnamese question:**
  "Dân số của Việt Nam năm 2021 có cao hơn khoảng gấp rưỡi so với Thái Lan không?"
  English translation:
  "Is the population of Viet Nam in 2021 approximately one and a half times larger than that of Thailand?"
- **Resolution:** The wording is normalised while preserving the underlying comparison task; the answer is checked and adjusted to match the actual ratio in the table.

B.3 Computational Errors (Corrected by Manual Recalculation)

Example 3: Incorrect percentage growth.

- **Underlying chart:** Growth rate (in %) of orange and mandarin production from 2017 to 2021.
- **Original Vietnamese question (generated by ChatGPT):**
  "Năm 2021, tỉ lệ tăng trưởng sản lượng cam, quýt là bao nhiêu?"
  English translation:
  "In 2021, what is the growth rate of orange and mandarin production?"
- **Original Vietnamese answer (generated by ChatGPT, incorrect):**
  "236,2%."
  English translation:
  "236.2%."
- **Issue:** The question is aligned with the chart, but 236.2% does not match the value obtained when the growth rate is computed from the data using the standard formula:

$$\text{growth rate} = \frac{\text{value}_{2021} - \text{value}_{2020}}{\text{value}_{2020}} \times 100.$$

- **Resolution:** The annotator manually recomputes the percentage based on the 2020 and 2021 values, replaces 236.2% with the correct value, and updates the Vietnamese instruction to describe the correct computation steps. The triplet is kept in the final dataset after correction.

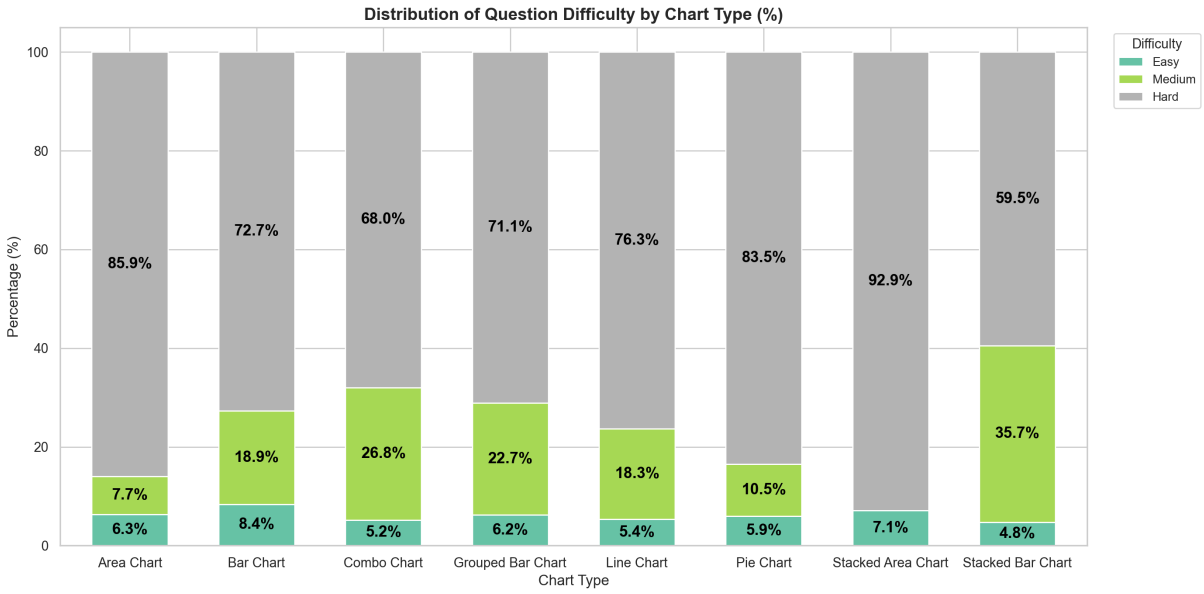


Figure 3: Distribution of question difficulty by chart type in ChartTableViVQA (%).

C Question difficulty distribution by chart types

Figure 3 shows the difficulty distribution broken down by chart types, including Area Chart, Bar Chart, Combo Chart, Grouped Bar Chart, Line Chart, Pie Chart, Stacked Area Chart, and Stacked Bar Chart.

D Implementation details

All experiments are implemented in PyTorch using official or community-maintained inference libraries. Inference is performed on a single NVIDIA RTX 3090 GPU via vast.ai.

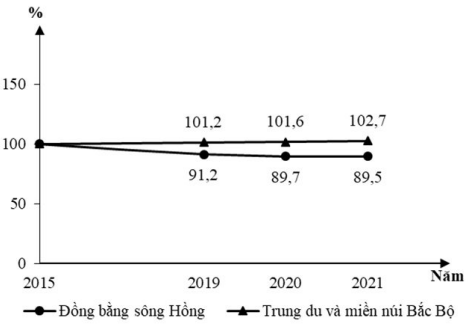
We use a unified decoding configuration for all models: maximum generation length of 150 tokens, beam size equals 7, temperature equals 0.4, top-p equals 0.9, and repetition penalty equals 1.5. Prompts instruct models to answer concisely in Vietnamese and include numerical values with units.

Before evaluation, model outputs undergo minimal normalisation (whitespace and punctuation), while all numerical values and units are preserved. The same post-processing is applied across models and settings to ensure fair comparison.

E Error predicted sample

Qualitative examples of numerical, mismatch, and comparison errors made by Qwen2.5-VL-7B-Instruct on ChartTableViVQA.

Comparison error



TỐC ĐỘ TĂNG TRƯỞNG SẢN LƯỢNG LÚA CỦA TRUNG DU VÀ MIỀN NÚI BẮC BỘ VÀ ĐỒNG BẰNG SÔNG HỒNG, GIAI ĐOẠN 2015 - 2021

(Số liệu theo Niên giám thống kê Việt Nam 2022, https://www.gso.gov.vn)

Question: Trong năm 2019, tốc độ tăng trưởng sản lượng lúa của Trung du và miền núi Bắc Bộ cao hơn hay thấp hơn Đồng bằng sông Hồng bao nhiêu phần trăm?
[English] In 2019, by how many percent was the growth rate of rice output in the Midland and Northern Mountainous region higher or lower than that of the Red River Delta?
Gold answer: Cao hơn 10%. **[English]** Higher by 10%.
Prediction: Thấp hơn 10%. **[English]** Lower by 10%.

Mismatch error

Cho bảng số liệu:
LAO ĐỘNG TỪ 15 TUỔI TRỞ LÊN PHÂN THEO THÀNH THỊ VÀ NÔNG THÔN CỦA NƯỚC TA, GIAI ĐOẠN 2015 - 2021

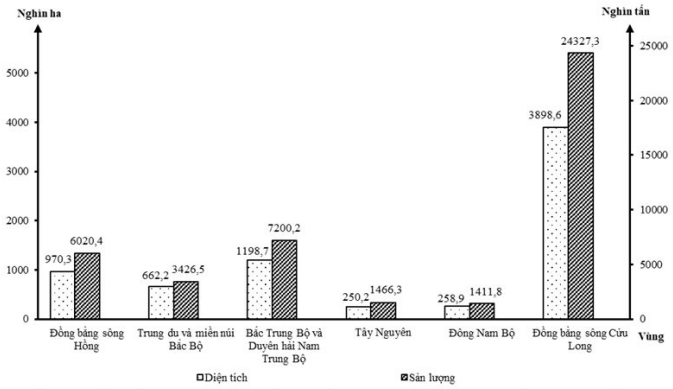
(Đơn vị: Nghìn người)

Năm	2015	2017	2020	2021
Thành thị	16913,8	17416,1	18171,93	18535
Nông thôn	37352,2	37403,5	36671,01	32025,5

(Nguồn: Niên giám thống kê Việt Nam 2021, NXB Thống kê, 2022)

Question: Lao động nông thôn giảm bao nhiêu phần trăm từ năm 2017 đến 2021?
[English] By how many percent did rural employment decrease from 2017 to 2021?
Gold answer: Giảm 14,38%. **[English]** Decreased by 14.38%.
Prediction: -14,38%. **[English]** -14.38%.

Numerical error



DIỆN TÍCH VÀ SẢN LƯỢNG LÚA CỦA CÁC VÙNG NƯỚC TA, NĂM 2021

(Số liệu theo Niên giám thống kê Việt Nam 2022, https://www.gso.gov.vn)

Question: Đông Nam Bộ trong năm 2021 sản xuất được bao nhiêu tấn lúa?
[English] How many tons of rice were produced in the Southeast region in 2021?
Gold answer: 1411,8 nghìn tấn. **[English]** 1411.8 thousand tons.
Prediction: 1411,8 tấn. **[English]** 1411.8 tons.

F Appendix: Examples of questions, answers, and instructions in Vietnamese and English in the ChartTableViVQA dataset

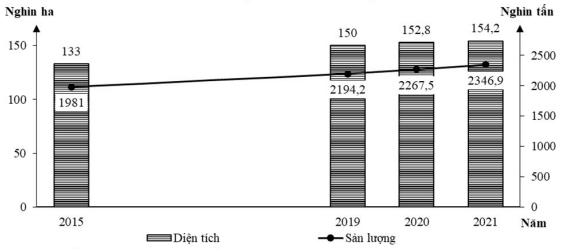
---

ĐẦU TƯ TRỰC TIẾP CỦA NƯỚC NGOÀI ĐƯỢC CẤP GIẤY PHÉP Ở VIỆT NAM

Năm	**Số dự án**	**Vốn đăng kí** *(triệu USD)*	**Vốn thực hiện** *(triệu USD)*
1991	152	1292	329
1995	415	6937	2556
1996	372	10164	2714
1997	349	5591	3115
2000	391	2839	2414
2005	970	6840	3309
2006	987	12004	4100

Question: Bảng trên thể hiện nội dung gì? [English: What does the table show?]
Answer: Bảng trên thể hiện đầu tư trực tiếp của nước ngoài được cấp giấy phép ở Việt Nam. [English: The table shows foreign direct investment licensed in Vietnam.]
Instruction: Cho bảng số liệu đầu tư trực tiếp của nước ngoài được cấp giấy phép ở Việt Nam. [English: Given the statistical table of foreign direct investment licensed in Vietnam.]
Question type / Difficulty: Image Info – Easy.

Cho biểu đồ về diện tích và sản lượng chuối của nước ta, giai đoạn 2015 - 2021:



(Số liệu theo Niên giám thống kê Việt Nam 2022, https://www.gso.gov.vn)

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 type / Difficulty: Lookup – Easy.

SẢN LƯỢNG VÀ CƠ CẤU GIÁ TRỊ SẢN XUẤT THỦY SẢN CỦA NƯỚC TA QUA CÁC NĂM

Năm / Sản lượng và Cơ cấu giá trị sản xuất	2000	2005	2010	2013
Sản lượng (nghìn tấn)	2 250,5	3 465,9	5 142,7	6 019,7
- khai thác	1 660,9	1 987,9	2 414,4	2 803,8
- nuôi trồng	589,6	1 478,0	2 728,3	3 215,9

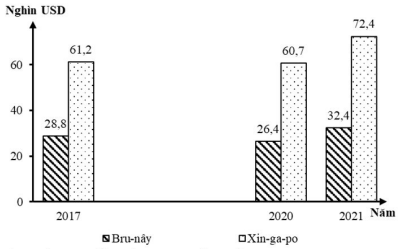
Question: Chênh lệch sản lượng nuôi trồng thuỷ sản giữa năm 2013 và năm 2005 là bao nhiêu nghìn tấn? [English: What is the difference in aquaculture production between 2013 and 2005 in thousand tonnes?]
Answer: 1737.9 nghìn tấn [English: 1737.9 thousand tonnes]
Instruction: Lấy giá trị hàng *Nuôi trồng* 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 for the *2005* column which is 1478.0: 3215.9 − 1478.0 = 1737.9.]
Question type / Difficulty: Aggregation – Medium.

DIỆN TÍCH GIEO TRỒNG CÂY CÔNG NGHIỆP LÂU NĂM NĂM 2013 CỦA CẢ NƯỚC, TRUNG DU VÀ MIỀN NÚI BẮC BỘ, TÂY NGUYÊN

(đơn vị : nghìn ha)

Cây công nghiệp	Cả nước	Trung du và miền núi Bắc	Tây Nguyên
Cây công nghiệp lâu	2134,9	142,4	969,0
Cà phê	641,2	15,5	573,4
Chè	132,6	96,9	22,9
Cao su	978,9	30,0	259,0
Các cây khác	382,2	0	113,7

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 năm*, cột *Cả nước (nghìn ha)* năm 2013 là 2134.9 và năm 2005 là 1633.6: $2134.9 > 1633.6$, và $2134.9 - 1633.6 = 501.3$ (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: $2134.9 > 1633.6$, and $2134.9 - 1633.6 = 501.3$ (higher).]
Question type / Difficulty: Comparison – Hard.



GDP BÌNH QUÂN ĐẦU NGƯỜI CỦA BRU-NÂY VÀ XIN-GA-PO, GIAI ĐOẠN 2017 - 2021

(Số liệu theo Niên giám thống kê ASEAN 2022, https://www.aseanstats.org)

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: $((32.4 - 26.4)/26.4) \times 100 = 22.73$ (tăng). [English: Compute the percentage change by taking (the value in the *Brunei (thousand USD)* column for 2021, which is 32.4, minus the value for 2020, which is 26.4), dividing by the value for 2020 (26.4), and multiplying by 100: $((32.4 - 26.4)/26.4) \times 100 = 22.73$ (increased).]
Question type / Difficulty: Trend Analysis – Hard.

G Appendix: Example of Chart Description in Both Image and JSON Data Structures

G.1 Chart Data Structures in JSON

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"
```

```
    }
    ...
  ]
}
```

G.2 Chart Image Example

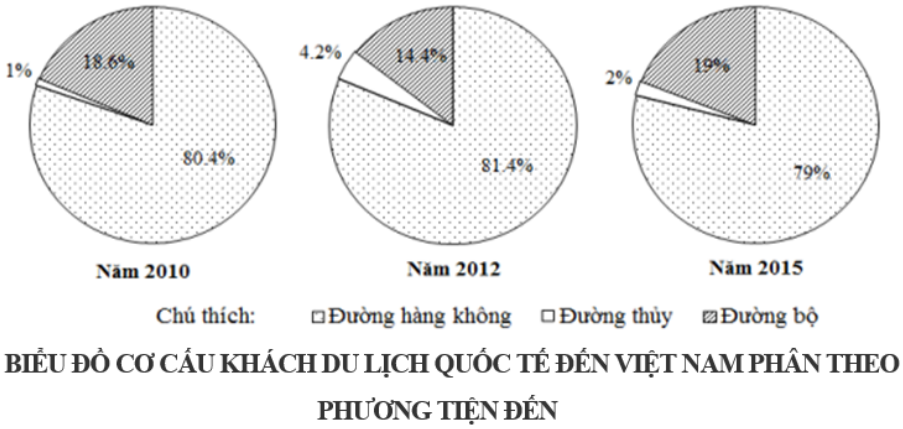


Figure 4: Pie chart showing the proportion of international tourists arriving in Vietnam by mode of transportation (2010, 2012, 2015) document ID 0870.

H Appendix: Example of Table Description in Both Image and JSON Data Structures

H.1 Table Data Structures in JSON

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.

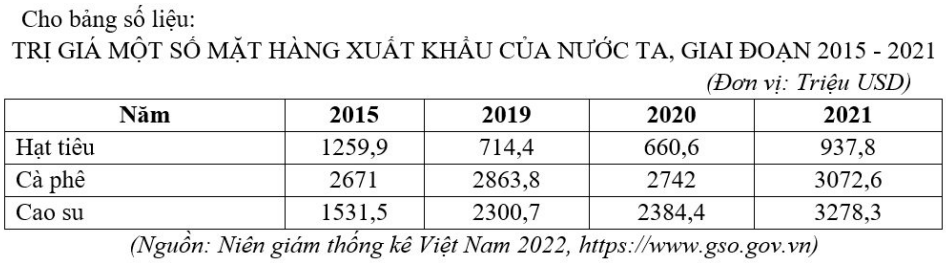
```
{
  "document_id": "0405",
  "image_info": {
    "image_name": "0405.jpg",
    "caption": "Trị giá một số mặt hàng xuất khẩu của nước ta, giai đoạn 2015 - 2021"
  },
  "columns": ["Năm", "2015", "2019", "2020", "2021"],
  "rows": [
    { "Năm": "Hạt tiêu", "2015": 1259.9, "2019": 714.4, "2020": 660.6, "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." }
  ]
}
```


H.2 Table Image Example



Cho bảng số liệu:

TRỊ GIÁ MỘT SỐ MẶT HÀNG XUẤT KHẨU CỦA NƯỚC TA, GIAI ĐOẠN 2015 - 2021

(Đơn vị: Triệu USD)

Năm	2015	2019	2020	2021
Hạt tiêu	1259,9	714,4	660,6	937,8
Cà phê	2671	2863,8	2742	3072,6
Cao su	1531,5	2300,7	2384,4	3278,3

(Nguồn: Niên giám thống kê Việt Nam 2022, https://www.gso.gov.vn)

Figure 5: Table showing the export values of selected commodities in Vietnam (2015-2021) document ID 0405.