INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Applying Multimodal Large Language Models for Visual Question Answering: Toward Vietnamese Educational Reasoning Systems

Xinh Le[1,2], Tho Quan[1,2*]
[1]Faculty of Computer Science and Engineering, Ho Chi Minh City University of Technology (HCMUT), 268 Ly Thuong Kiet Street, Dien Hong Ward, Ho Chi Minh City, Vietnam
[2]Vietnam National University Ho Chi Minh City, Linh Xuan Ward, Ho Chi Minh City, Vietnam
Email: ltxinh.sdh251@hcmut.edu.vn; qttho@hcmut.edu.vn*

Tien-Thinh Nguyen[3,4]
[3]Faculty of Information Technology
[4]Industrial University of Ho Chi Minh City
Ho Chi Minh City, Vietnam
Email: 01036040_thinh@iuh.edu.vn

Submitted: 30/9/2025. Revised edition: 14/4/2026. Accepted: 4/5/2026. Published online: 10/6/2026
DOI: https://doi.org/10.11113/ijic.v16n1.680

***Abstract*—Visual Question Answering (VQA) is rapidly advancing due to Multimodal Large Language Models (MLLMs), which demonstrate powerful complex reasoning capabilities. However, this progress is predominantly centered on English and general-domain contexts. Domain-specific fields, such as STEM education, and low-resource languages, such as Vietnamese, remain significantly underserved, lacking both standardized datasets and specialized reasoning models. This research addresses this gap by investigating how MLLMs can be adapted for Vietnamese educational settings. As a foundational step, the ViPPS dataset (Vietnamese Physics Problem Solving) has been constructed and publicly released, the first multimodal dataset for physics problem solving in Vietnamese. Initial experiments on ViPPS show that current MLLMs achieve promising results, but still struggle with domain-specific reasoning and numerical accuracy. Based on these observations, the next stage of this research will focus on expanding the ViPPS dataset in both scale and scope, developing and evaluating advanced text-image reasoning and calculation mechanisms and extending these capabilities to VideoQA and model explainability. This research will contribute critical resources and methods, advancing the field of educational VQA for low-resource languages.**

***Keywords*—Visual Question Answering (VQA), Multimodal Large Language Models (MLLMs), STEM Education, Visual Reasoning, Chain-of-Thought Reasoning**

I. INTRODUCTION

The recent emergence of Multimodal Large Language Models (MLLMs), such as LLaVA [1] and GPT-4V [2], has revolutionized the field of Visual Question Answering (VQA). These models integrate the powerful reasoning abilities of Large Language Models (LLMs) with visual comprehension, enabling them to understand and reason about complex visual and textual information simultaneously. Their success in general-domain benchmarks demonstrates a strong potential for applications requiring a nuanced understanding of the world.

Despite this progress, two significant gaps persist. First, current state-of-the-art models are trained predominantly on general-domain English data. This focus limits their ability to reason in specialized domains like STEM education, which often requires deep subject knowledge (e.g., physics principles, chemical laws) and precise numerical calculation - abilities that general-domain models inherently lack. This problem is exacerbated for low-resource languages like Vietnamese. The Vietnamese NLP landscape suffers from a severe scarcity of large-scale, high-quality multimodal datasets. This lack of resources hinders the development of AI-powered educational tools that could significantly benefit Vietnamese students, creating a digital divide in advanced AI applications.

Motivated by this urgent need, this research is dedicated to building robust MLLM-based reasoning systems specifically tailored for the Vietnamese educational context. The goal is to move beyond simple image captioning or basic VQA, and develop systems that can reason, calculate, and explain complex STEM problems, acting as a valuable assistant for students and educators. In practice, such systems are expected to function as intelligent tutoring assistants that guide students through step-

by-step problem solving, providing explanations and intermediate reasoning rather than only final answers. In addition, they may support automated assessment by evaluating student responses and identifying common reasoning errors. The primary research objectives of this dissertation are: (1) To establish a foundational benchmark dataset for Vietnamese educational VQA, an objective now complete via the ViPPS dataset paper [3]; (2) To expand this data resource in both scale and scope to other STEM subjects; (3) To design, implement, and evaluate novel MLLM mechanisms capable of robust logical reasoning and numerical calculation; and (4) To extend these capabilities from static images to dynamic videos (VideoQA) and analyze model explainability.

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of related work. Section 3 describes the ViPPS dataset [3] and presents preliminary evaluation results, with a focus on the main limitations revealed through benchmarking. Section 4 presents the three-pronged future research plan that forms the core of this dissertation. Finally, section 5 concludes the paper with a summary of contributions and potential impact.

II. RELATED WORKS

A. Architectures of Multimodal Large Language Models

The dominant paradigm in MLLMs involves connecting a pre-trained visual encoder with a pre-trained LLM. This architecture typically consists of three components: a vision encoder (e.g., CLIP-ViT [4]) that extracts visual features, an LLM (e.g., LLaMA [5]) that serves as the reasoning backbone, and an adaptation module that maps visual features into the LLM's embedding space. The design of this adaptation module varies, LLaVA [1] employs a simple Multi-Layer Perceptron (MLP), whereas models like BLIP-2 [6] and MiniGPT-4 [7] utilize a more complex cross-modal transformer module to distill visual features into a fixed number of learnable queries.

The training process for these models is generally two-staged. First, a vision-language pre-training phase aligns the vision and language modalities, often using massive, web-scale image-text pair datasets. This is followed by a visual instruction tuning phase, where the model is fine-tuned on a smaller, high-quality dataset of multimodal instructions (e.g., VQA pairs, complex reasoning tasks) to elicit the desired conversational and reasoning behaviors. LLaVA, for instance, pioneered the use of GPT-4 [8] to generate this instruction data, demonstrating that the quality of this tuning data is critical for strong performance.

B. Benchmarks for Educational and STEM VQA

While general-domain VQA benchmarks like VQAv2 [9] have driven progress, they often test descriptive understanding rather than deep reasoning. In response, specialized benchmarks for STEM have emerged. ScienceQA [10] is a prominent example, providing questions that require understanding a multimodal context (an image, a text snippet, or both). It tests a wide range of skills, including visual grounding, text retrieval, and multi-step reasoning. MathQA [11] focuses on mathematical problems presented in images, requiring robust Optical Character Recognition (OCR) to read the problem text and figures, followed by mathematical reasoning to derive the solution.

However, these benchmarks have several limitations that are particularly relevant to this work. First, they are overwhelmingly English-centric, offering no resources for low-resource languages. Second, many (though not all) are formatted as multiple-choice questions, which tests discriminative ability but does not evaluate a model's capacity to generate free-form, step-by-step explanations - a crucial function for an educational tool.

The Vietnamese AI landscape has seen progress in general-domain VQA. OpenViVQA [12] and ViTextVQA [13] are two large-scale datasets. However, their focus is on general scene understanding and scene-text comprehension, respectively. They do not contain the domain-specific knowledge, diagrams (like circuits or mechanics), or multi-step reasoning problems characteristic of STEM education. While strong Vietnamese LLMs like PhoBERT [14] and ViT5 [15] exist, they lack the visual grounding necessary to tackle these multimodal tasks. This creates a clear research gap: a lack of both datasets and models for specialized, multimodal STEM reasoning in the Vietnamese educational context. The ViPPS dataset [3] was developed as an initial step toward addressing this gap.

III. FOUNDATIONAL WORK: THE VIPPS DATASET

This section presents the completed components of this research, focusing on the construction of the ViPPS dataset and the initial benchmarking of existing Vision–Language Models (VLMs). These results establish an empirical foundation for understanding the current capabilities and limitations of multimodal reasoning in Vietnamese educational contexts. The findings reported here are based on previously published work [3] and serve as the basis for identifying open challenges, which are addressed in the subsequent section as part of the proposed future research directions.

A. Dataset Construction Pipeline

This detailed rationale is a key feature of the dataset, designed explicitly to support the training of reasoning models. The ViPPS dataset was constructed through a four-stage pipeline (Figure 1), with each stage designed to progressively improve data quality and consistency [3].

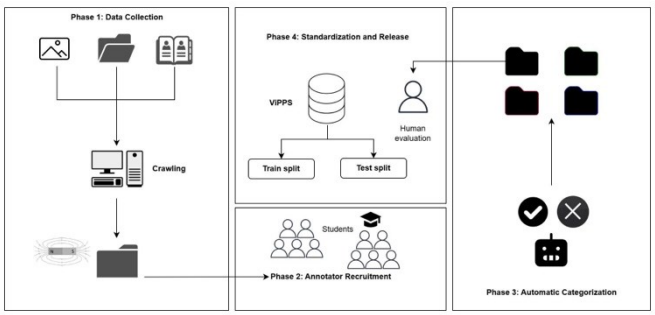

Fig. 1. Overview of the ViPPS dataset construction process

Phase 1. Data collection: Physics problems and their corresponding diagrams were collected from multiple sources, including Vietnamese educational Q&A platforms, online textbooks, and exam preparation materials. This process aimed to capture a broad spectrum of visual formats, ranging from well-structured schematic diagrams to informal hand-drawn sketches shared by students, thereby reflecting realistic learning scenarios.

Phase 2. Annotator selection: A group of 10 annotators was recruited, consisting of Grade 12 students with strong backgrounds in physics. To maintain annotation quality, a predefined evaluation rubric was applied during the selection process, focusing on correctness in problem understanding, consistency in reasoning, and clarity in explanation. Only candidates who satisfied these criteria were selected, ensuring that the annotation process was carried out by individuals with sufficient domain knowledge and reliable interpretation skills.

Phase 3. Automatic categorization: To facilitate large-scale processing, candidate questions were assigned to predefined physics categories using multiple large language models. Only instances with full agreement across all three models were directly accepted. Cases with partial agreement were retained but flagged for further inspection. This consensus-based filtering serves as an initial mechanism to improve labeling consistency before manual verification.

Phase 4. Standardization and release: All flagged instances were subsequently reviewed by expert annotators to correct misclassifications, resolve ambiguities, and ensure proper alignment between problem statements and their associated diagrams. This stage also functions as a final quality control step. Although an explicit inter-annotator agreement (IAA) score is not reported, multiple measures were applied to improve annotation reliability. In particular, only samples that remained consistent after both automatic filtering and manual verification were retained. This multi-stage validation process helps reduce annotation inconsistencies and improve overall data quality. A formal quantitative assessment of annotation agreement will be considered in future work.

The final release of ViPPS contains 5,520 physics problems, including 3,750 multimodal instances where each question is associated with an image, and 1,770 text-only questions. These multimodal samples form image–question–answer triplets that serve as the core component of the dataset.

Each instance is organized in a consistent format, including the image (if available), the question, the corresponding answer, a short caption describing the visual content, and additional metadata such as problem ID and topic category. This structure allows the dataset to be used flexibly for different types of multimodal reasoning tasks.

In total, the dataset includes 448 unique images. These images vary considerably in style, ranging from clean textbook diagrams to hand-drawn sketches and photos shared by students. This variation reflects how physics problems are encountered in practice and introduces different levels of visual difficulty for model evaluation.

Overall, ViPPS provides a balanced coverage of key domains in high school physics, including electricity, magnetism, and electromagnetism, with a strong emphasis on circuit-based reasoning. The dataset reflects authentic learning con- texts by combining formal textbook-style exercises with real student-generated problem statements and diagrams.

B. ViPPS Evaluation and Key Findings

To establish an empirical baseline for multimodal reasoning in Vietnamese educational settings, an evaluation framework is designed that captures not only answer correctness but also model reliability and robustness. Specifically, accuracy is used as the primary metric to measure performance across different question types. In addition, calibration quality is assessed using Expected Calibration Error (ECE) and Brier Score, which reflect how well model confidence aligns with actual correctness. This aspect is particularly important in educational applications, where unreliable confidence estimates may negatively affect learners. Accuracy@Coverage is used to assess selective prediction behavior, focusing on performance when only high-confidence predictions are retained. Finally, robustness is evaluated under controlled visual perturbations, including noise injection, scaling, rotation, and symbol variations, to simulate realistic distortions commonly found in diagrams and handwritten inputs.

The detailed experimental results and full analysis are reported in ViPPS dataset paper [3]. Here, the experimental setup is summarized and key observations relevant to this study.

To ensure transparency and reproducibility, the baseline model configurations are explicitly defined. A representative set of models is evaluated, including Qwen2-VL-7B, LLaVA-1.6-13B, InternVL2-8B, IDEFICS2-8B, InstructBLIP-7B, mPLUG-Owl2-7B, and MiniGPT-4 (7B). For reference, results are also included from larger proprietary systems such as GPT-4V and Gemini PV. This selection covers both open-source and commercial models, allowing for a broad comparison across architectures and model scales.

Each model is evaluated under three experimental settings. In the zero-shot setting, models directly generate answers from the input image and question without any additional adaptation. In the caption-augmented setting (I+Cap+Q), an additional textual description of the image is provided alongside the input, serving as an intermediate semantic representation to support reasoning. In the fine-tuning setting, models are trained on the ViPPS training split, with and without caption augmentation, enabling adaptation to the characteristics of Vietnamese physics problems. Unless otherwise specified, default inference configurations provided by each model are used, and no task-specific prompt engineering is introduced beyond standard input formatting.

Across these configurations, several consistent patterns emerge. First, incorporating captions into the input generally improves performance in the zero-shot setting, suggesting that textual descriptions help bridge the gap between visual perception and language reasoning. Second, fine-tuning yields the most substantial performance gains, improving not only overall accuracy but also confidence calibration. When combined, caption augmentation and fine-tuning produce more stable and reliable predictions, indicating a complementary effect between explicit visual descriptions and task-specific adaptation.

Robustness analysis further shows that fine-tuned models maintain relatively stable performance under moderate visual perturbations, although degradation becomes more noticeable as distortions increase. This suggests that domain adaptation contributes to improved resilience against variations in diagram quality and representation.

Despite these improvements, several limitations remain evident. Current VLMs still struggle with problems that require multi-step reasoning or precise numerical computation, particularly in more complex physics domains or when dealing with informal, hand-drawn diagrams. These challenges highlight the need for more structured reasoning mechanisms and tighter integration with external computational tools.

Overall, the ViPPS benchmark provides a useful diagnostic for evaluating multimodal reasoning in low-resource educational contexts. While existing models demonstrate promising capabilities, the results indicate that further progress will depend on domain-specific training, improved reasoning strategies, and better calibration of model confidence.

IV. PROPOSED FUTURE RESEARCH

Building on the completed results presented in Section III, the remainder of this research focuses on addressing the identified limitations through three main directions.

A. Dataset Expansion: Beyond Physics

Building upon the foundational ViPPS dataset, the first contribution addresses its current limitations in both scale and domain coverage. While ViPPS provides valuable insight into multimodal reasoning for physics education, it remains restricted to a single subject and limited dataset size, thereby constraining model generalization and diversity of reasoning patterns. This leads to the following research questions:

(RQ1.1) How can data collection be efficiently scaled beyond conventional scraping-based methods?

(RQ1.2) What new visual reasoning challenges are introduced by other STEM domains like Chemistry and Biology?

To address these questions, the ViPPS dataset will be extended in two main directions: increasing its scale and expanding its coverage to other domains. The first direction focuses on improving the data collection pipeline by incorporating semi-automated retrieval and data augmentation techniques. These methods enable controlled variation of problem parameters and the generation of corresponding answers in a consistent manner. The second direction involves the construction of additional multimodal datasets for Chemistry and Biology. These domains introduce different types of reasoning, such as interpreting molecular structures, analyzing reaction diagrams, and understanding cellular or anatomical processes.

The expected outcome of this effort is ViM-STEM (Vietnamese Multimodal STEM), a comprehensive dataset encompassing multiple STEM disciplines. ViM-STEM will represent a major step toward multimodal research for low-resource languages, offering a benchmark for evaluating visual reasoning, scientific understanding, and text–vision alignment in Vietnamese educational contexts.

B. Developing Advanced Reasoning and Calculation Mechanisms

This contribution directly tackles the core limitations of reasoning and calculation. The second contribution focuses on improving reasoning transparency and computational reliability, two key weaknesses identified in current MLLMs. Despite their linguistic fluency, these models often fail in logical consistency and numerical accuracy, which limits their trustworthiness in educational problem-solving. This leads to two guiding questions:

(RQ2.1) Can explicitly fine-tuning on step-by-step rationales (Chain-of-Thought) improve a model's logical consistency and faithfulness?

(RQ2.2) How can symbolic solvers be effectively integrated to address the numerical limitations of MLLMs?

To investigate these questions, two complementary directions are considered. The first focuses on rationale-based fine-tuning, where existing step-by-step solutions in ViPPS are reformulated into structured reasoning traces. This allows the model to generate intermediate reasoning steps instead of directly predicting final answers. The second direction explores tool-augmented MLLMs, in which computational steps are delegated to symbolic solvers. This separation enables the model to concentrate on reasoning, while numerical results are verified through external computation.

The expected outcome is a reasoning-augmented framework that integrates external tools to improve both interpretability and numerical accuracy. Such a framework supports more reliable and transparent problem solving in educational settings.

C. Extension to Lecture VideoQA and Explainability

The third direction extends the scope of this research from static multimodal problems to lecture-based settings, where information is distributed across time. Unlike images or diagrams, lecture videos involve two complementary yet not always synchronized modalities: visual content (e.g., slides, figures, on-screen text) and spoken explanations. In practice, a single slide may be discussed over an extended period, creating a complex relationship between what is presented visually and what is explained verbally. This temporal mismatch makes multimodal reasoning more challenging.

Based on this setting, two research questions are considered:

(RQ3.1) How can spoken transcripts be aligned with the corresponding visual content in lecture slides?

(RQ3.2) What mechanisms allow models to utilize this alignment for accurate and interpretable multimodal reasoning?

To address these questions, a new benchmark dataset, ViE-LecQA (Vietnamese Educational Lecture QA), will be constructed from Vietnamese lecture videos. The process includes speech transcription and slide content extraction, followed by manual design of questions that require integrating information across modalities. For example, a formula may

appear on a slide while specific values are introduced only in the spoken explanation.

Based on this dataset, an Aligned LectureQA framework is developed to first associate transcript segments with relevant slide regions, and then perform reasoning over the aligned multimodal inputs.

The expected outcome is a Vietnamese benchmark and modeling approach for lecture-based VQA. By producing grounded explanations that explicitly reference both visual and temporal evidence (e.g., slide regions and timestamps), the system can improve interpretability and support practical educational applications such as tutoring and assessment.

V. CONCLUSION

This research investigates the use of MLLMs for developing educational reasoning systems for Vietnamese. A foundational result has already been established by creating and benchmarking the ViPPS dataset [3]. This initial work was critical in confirming the potential of MLLMs for this task, but more importantly, in empirically identifying their core weaknesses, a failure in deep logical reasoning and an inability to perform precise numerical calculations.

The development and benchmarking of the ViPPS dataset demonstrate the feasibility and potential of MLLMs in handling multimodal educational tasks. At the same time, the initial findings reveal persistent challenges, particularly in deep logical reasoning and numerical computation, which remain critical barriers to deploying MLLMs in authentic learning environments.

To address these limitations, the research is organized into three main directions. The first focuses on expanding the dataset in both scale and domain coverage, extending ViPPS toward a broader ViM-STEM resource that includes multimodal learning contexts in Physics, Chemistry, and Biology. The second direction aims to improve reasoning and computational performance by incorporating structured reasoning strategies, such as chain-of-thought, together with external tool integration for numerical verification. The third direction explores extensions to VideoQA and explainable AI, with the goal of capturing temporal reasoning in lecture-based settings and improving the interpretability of model outputs.

If successful, this research will deliver the first large-scale multimodal STEM dataset in Vietnamese, alongside a new generation of MLLMs capable of genuine reasoning, quantitative problem-solving, and explainable decision-making. Beyond technical innovation, the expected outcomes have direct implications for AI-powered tutoring, assessment automation, and personalized learning, contributing to equitable access to high-quality education for millions of Vietnamese students. Furthermore, the findings are anticipated to enrich the global understanding of domain-specific multimodal reasoning, positioning Vietnamese as an emerging testbed for low-resource AI research.

ACKNOWLEDGMENTS

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM for supporting this study.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024)* (pp. 26296–26306). https://doi.org/10.1109/CVPR52733.2024.02484.

[2] Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., Liu, B., & Tao, J. (2024). GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, *108*. https://doi.org/10.1016/j.inffus.2024.102367.

[3] Vo, Q. T. N., Le, X. T., Tran, T. H. M., & Quan, T. T. (2025). ViPPS: Building a multimodal dataset for physics problem solving in Vietnamese. In *Proceedings of the 18th Multi-Disciplinary International Conference on Artificial Intelligence (MIWAI 2025)* (pp. 308–319). Springer. https://doi.org/10.1007/978-981-95-4960-3_25

[4] Larkin, V. D., Ivanov, Y. S., & Chukhnov, A. P. (2025). One-shot visual detection of phishing resources with CLIP ViT and contrastive learning. In *Proceedings of the 2025 International Russian Smart Industry Conference (SmartIndustryCon)* (pp. 837–842). IEEE. https://doi.org/10.1109/SmartIndustryCon65166.2025.10986239

[5] Caumartin, G., Qin, Q., Chatragadda, S., Panjrolia, J., Li, H., & Costa, D. E. (2025). Exploring the potential of Llama models in automated code refinement: A replication study. In *Proceedings of the 2025 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)* (pp. 681–692). IEEE. https://doi.org/10.1109/SANER64311.2025.00070.

[6] Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)* (pp. 19730–19742). https://doi.org/10.5555/3618408.3619222

[7] Tzelepi, M., & Mezaris, V. (2024). Exploiting LMM-based knowledge for image classification tasks. In *Proceedings of the 25th International Conference on Engineering Applications of Neural Networks (EANN 2024)* (pp. 166–177). Springer. https://doi.org/10.1007/978-3-031-62495-7_13

[8] Tait, I., Bensemann, J., & Wang, Z. (2024). Is GPT-4 conscious? *Journal of Artificial Intelligence and Consciousness*, *11*(1), 1–16. https://doi.org/10.1142/S270507852450005X.

[9] Long, Y., Tang, P., Wang, H., & Yu, J. (2021). Improving reasoning with contrastive visual information for visual question answering. *Electronics Letters*, *57*(20), 758–760. https://doi.org/10.1049/ell2.12255.

[10] Qiu, C., Xie, Z., Liu, M., & Hu, H. (2024). Explainable knowledge reasoning via thought chains for knowledge-based visual question answering. *Information Processing & Management*, *61*, 103726. https://doi.org/10.1016/j.ipm.2024.103726.

[11] Amini, A., Gabriel, S., Lin, S., Kedziorski, R. K., Choi, Y., & Hajishirzi, H. (2019). MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019)* (pp. 2357–2367). https://doi.org/10.18653/v1/N19-1245.

[12] Nguyen, N. H., Vo, D. T. D., Nguyen, K. V., & Nguyen, N. L. T. (2023). OpenViVQA: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese. *Information Fusion*, *100*, 101868. https://doi.org/10.1016/j.inffus.2023.101868.

[13] Nguyen, Q. V., Tran, D. Q., Pham, H. Q., Nguyen, T. K. B., Nguyen, N. H., Nguyen, K. V., & Nguyen, N. L. T. (2026). ViTextVQA: A large-scale visual question answering dataset and a novel multimodal feature fusion method for Vietnamese text comprehension in images. *Expert Systems with Applications*, *308*, 130839. https://doi.org/10.1016/j.eswa.2025.130839.

[14] Nguyen, H. T., Huynh, T. N., Mai, N. T. N., Le, K. D. D., & Pham, D. T. N. (2023). PhoBERT application in disease classification based on Vietnamese symptom analysis. *Applied Computer Systems*, *28*(1), 35–43. https://doi.org/10.2478/acss-2023-0004.

[15] Phan, L., Tran, H., Nguyen, H., & Trieu, T. H. (2022). ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop (NAACL 2022)* (pp. 136–142). https://doi.org/10.18653/v1/2022.naacl-srw.18.