

Vietnamese Retrieval-Augmented Vision-Language Model (vRA-VLM)

Huong Le Thanh¹ Dung Vu Minh² Thuan Tran Xuan²

¹ Hanoi University of Science and Technology ² Hanoi University of Science and Technology
huong.lt@soict.hust.edu.vn, dung.vm205179@sis.hust.edu.vn, thuan.tx210824@sis.hust.edu.vn

Abstract

The development of Multi-Modal Large Language Models (MLLMs) has enabled the integration of various data modalities, such as images and text. However, conventional MLLMs frequently suffer from hallucinations, generating information that is unsupported by the input. In this work, we present the **Vietnamese Retrieval-Augmented Vision Language Model (vRA-VLM)**, which enhances multimodal understanding by incorporating an external knowledge retrieval mechanism. We describe the architecture of vRA-VLM, the retrieval strategies employed, and present comprehensive experiments on Vietnamese visual-language tasks. Results show that vRA-VLM outperforms baseline MLLMs by significantly reducing hallucination and improving task performance. All code and model weights are available at: <https://github.com/VuSnow/Vi-VLM-TTDDN.git>

1. Introduction

In recent years, Multi-Modal Large Language Models (MLLMs) have achieved significant breakthroughs in artificial intelligence by enabling machines to process and reason over diverse data types, including text, images, and audio. These models demonstrate remarkable performance across various applications, such as visual question answering (VQA), image captioning, multi-modal retrieval, and knowledge extraction. Despite this progress, conventional MLLMs still suffer from hallucination—generating information that is inaccurate or unsupported by input data—which undermines model reliability, particularly in domains requiring factual consistency.

To address this, retrieval-augmented generation techniques have been introduced, leveraging external knowledge sources to ground model outputs in verifiable facts and reduce hallucination. While such approaches have proven effective, most existing research and benchmarks focus exclusively on English or other high-resource languages, leaving low-resource languages like Vietnamese largely underexplored.

Vietnamese presents unique linguistic challenges, and the scarcity of publicly available multi-modal datasets further complicates progress. The effectiveness of existing MLLMs on Vietnamese data is still unclear.

In this work, we introduce vRA-VLM—a novel retrieval-augmented vision-language model tailored for Vietnamese. Our architecture combines an external retrieval module, a powerful pretrained vision encoder, and a large Vietnamese language model, all integrated via an efficient cross-attention fusion mechanism. By conditioning model outputs on both visual and retrieved textual contexts, vRA-VLM aims to reduce hallucination and improve factual consistency. We extensively evaluate vRA-VLM on multiple Vietnamese multi-modal tasks, including VQA, image captioning, and document understanding, demonstrating substantial improvements over existing baselines.

2. Related work

Vision-Language Models (VLM): The past several years have witnessed rapid advancements in Vision-Language Models (VLMs), which aim to bridge the gap between visual and textual understanding for a variety of multi-modal tasks. Early approaches such as CLIP (1) leveraged contrastive learning to align image and text representations in a shared embedding space, achieving strong performance on zero-shot classification and retrieval.

Building on these foundations, more recent models have focused on enhancing the reasoning and generative capabilities of VLMs. BLIP (2) (Bootstrapped Language Image Pretraining) introduced a unified vision-language pretraining strategy by combining image-text contrastive learning with language modeling objectives, leading to improvements in both understanding and generation tasks. BLIP-2 (5) further advanced the field by decoupling the vision encoder and large language model, introducing a Querying Transformer (Q-Former) as an intermediary to efficiently connect frozen pretrained components. This architecture enables flexible integration of strong vision and language models, while significantly reducing the cost of multi-modal pretraining.

The introduction of cross-attention between image features and language tokens enables modern VLMs to achieve more granular and context-aware understanding of visual inputs, effectively overcoming the limitations of simple embedding alignment or concatenation. This mechanism allows each language token to dynamically attend to relevant visual regions, which is critical for tasks requiring precise grounding and factual consistency. As a result,

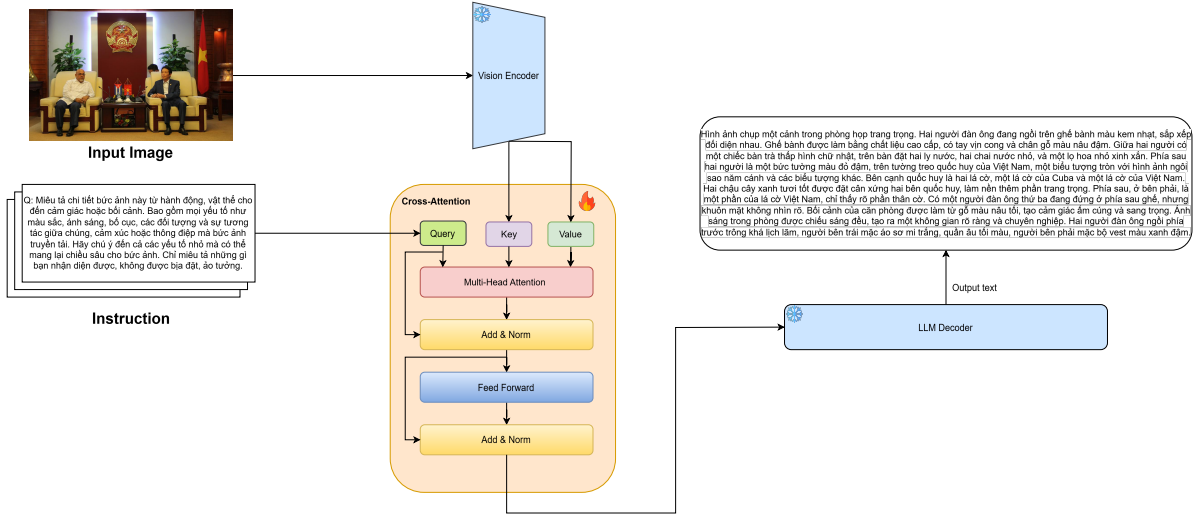


Figure 1: The overall architecture of stage 1 of the proposed vRA-VLM model.

cross-attention is a key factor behind the improved factual grounding and reduced hallucination observed in state-of-the-art models.

Flamingo (4) pushed multi-modal reasoning further by incorporating gated cross-attention layers between visual features and language tokens, demonstrating state-of-the-art results in few-shot visual question answering and captioning. Similarly, LLaVA (Large Language and Vision Assistant) (6) leveraged instruction tuning and multi-stage alignment between vision encoders and large language models, enabling open-ended visual reasoning capabilities that approach human-level performance in some benchmarks.

Despite their impressive performance, these modern VLMs still face persistent challenges, most notably hallucination—generating information unsupported by the provided visual or textual input. Additionally, the majority of these models are developed and evaluated primarily on English-centric datasets, limiting their applicability to low-resource languages such as Vietnamese.

Retrieval-Augmented Generation (RAG) for Large Language Models.: Traditional large language models (LLMs) generate responses based solely on their internal parameters and pretraining data, which limits their ability to provide up-to-date or highly factual information, and often leads to hallucination. To address this, retrieval-augmented

generation (RAG) frameworks have been proposed, which augment language models with an external retrieval component that grounds model outputs in relevant documents or knowledge bases.

A pioneering work in this direction is RAG (8), which integrates dense retrieval over a large corpus with a sequence-to-sequence language model, enabling dynamic conditioning of generated outputs on retrieved contexts. Subsequent advances, such as REALM (9) and RETRO (10), further explored end-to-end retrieval and retrieval-augmented pre-training at scale. These approaches have demonstrated substantial improvements in factual accuracy and knowledge-intensive tasks by reducing reliance on model memorization.

More recently, retrieval augmentation has been extended to multi-modal models. For example, BLIP-2 (5) and LLaVA-Retrieval incorporate retrieval mechanisms to supply additional textual or visual context, improve performance on vision-language tasks and mitigating hallucination. Retrieval-augmented models typically follow a two-stage pipeline: first, they retrieve relevant passages or facts from an external source based on the input query or image; then, the retrieved content is used as additional context for the language model to generate a grounded and informed response.

Despite these advances, the majority of retrieval-augmented research has focused on English or other

Input Image	Generated Description
	Bức ảnh chụp một gia đình Hàn Quốc, có vẻ là chụp ngoài trời, trước một tòa nhà màu kem với cửa sổ lớn. Bức ảnh có vẻ được chụp vào những năm 1960 hoặc 1970, được đánh giá qua kiểu quần áo và kiểu tóc. Có bốn người trong ảnh: một người đàn ông, hai phụ nữ lớn tuổi và một phụ nữ trẻ hơn. Người đàn ông ngồi ở giữa, mặc bộ vest màu xám đậm, cà vạt đỏ đậm, và áo sơ mi trắng. Ông ta ngồi trên ghế mây màu be, hai tay đặt trên đùi. Khuôn mặt ông ta nhìn thẳng vào ống kính, thể hiện vẻ nghiêm nghị nhưng thân thiện. Hai phụ nữ lớn tuổi đứng phía sau ông ta, trông có vẻ là vợ và con gái của ông. Người phụ nữ ở bên trái mặc áo khoác màu hồng nhạt trên váy, vẻ mặt trang nghiêm. Người phụ nữ ở bên phải mặc áo khoác họa tiết vàng và đen, cũng với vẻ nghiêm trang. Người phụ nữ trẻ hơn, có lẽ là con gái hoặc cháu gái, ngồi bên trái người đàn ông, mặc Hanbok (trang phục truyền thống Hàn Quốc) màu xanh nhạt với chiếc thắt lưng màu nâu đậm. Cô ấy có vẻ thanh lịch và điềm đạm. Bối cảnh bao gồm một khu vườn được cắt tỉa gọn gàng với những cây xanh và các chậu cây cảnh. Mặt trước của tòa nhà trông khá hiện đại và sạch sẽ, thể hiện sự giàu có và địa vị xã hội của gia đình.
	Hình ảnh chụp một cảnh trong phòng họp trang trọng. Hai người đàn ông đang ngồi trên ghế bành màu kem nhạt, sắp xếp đối diện nhau. Ghế bành được làm bằng chất liệu cao cấp, có tay vịn cong và chân gỗ màu nâu đậm. Giữa hai người có một chiếc bàn trà thấp hình chữ nhật, trên bàn đặt hai ly nước, hai chai nước nhỏ, và một lọ hoa nhỏ xinh xắn. Phía sau hai người là một bức tường màu đỏ đậm, trên tường treo quốc huy của Việt Nam, một biểu tượng tròn với hình ảnh ngôi sao năm cánh và các biểu tượng khác. Bên cạnh quốc huy là hai lá cờ, một lá cờ của Cuba và một lá cờ của Việt Nam. Hai chậu cây xanh tươi tốt được đặt cân xứng hai bên quốc huy, làm nền thêm phần trang trọng. Phía sau, ở bên phải, là một phần của lá cờ Việt Nam, chỉ thấy rõ phần thân cờ. Có một người đàn ông thứ ba đang đứng ở phía sau ghế, nhưng khuôn mặt không nhìn rõ. Bối cảnh của căn phòng được làm từ gỗ màu nâu tối, tạo cảm giác ấm cúng và sang trọng. Ánh sáng trong phòng được chiếu sáng đều, tạo ra một không gian rõ ràng và chuyên nghiệp. Hai người đàn ông ngồi phía trước trông khá lịch lãm, người bên trái mặc áo sơ mi trắng, quần âu tối màu, người bên phải mặc bộ vest màu xanh đậm.

Figure 2: Examples of input images and output descriptions from vRA-VLM.

high-resource languages. The application of RAG in low-resource languages such as Vietnamese remains limited, with unique challenges due to the scarcity of high-quality retrieval corpora and multi-modal datasets. Our work extends retrieval-augmented generation to Vietnamese vision-language modeling, leveraging both visual and external textual evidence to enhance factual grounding and reduce hallucination in multi-modal tasks.

Integrating Vision-Language Models and Retrieval Mechanisms.: Recent research has increasingly focused on the integration of vision language models (VLMs) with retrieval mechanisms to further enhance factual grounding and multi-modal understanding. Instead of relying solely on prelearned parameters, these architectures dynamically

incorporate external information relevant to a given input, enabling the model to access more precise and up-to-date knowledge.

- **Wiki-LLaVA (7):** LLaVA builds on the architecture with a hierarchical retrieval pipeline, sourcing relevant information from external knowledge bases such as Wikipedia. By integrating retrieved textual evidence with visual input, Wiki-LLaVA improves performance in visual question answering tasks that require external knowledge.
- **Re-ViLM (11):** Based on the Flamingo model, Re-ViLM introduces a retrieval-augmented framework for zero- and few-shot image captioning. By retrieving pertinent information from external databases, Re-ViLM enhances caption generation, particularly in out-


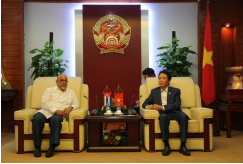
Images Input	Question	Answer
	Loại trang phục mà người phụ nữ trẻ tuổi đang mặc là gì và điều đó nói lên điều gì về bối cảnh hình ảnh?	Người phụ nữ trẻ tuổi mặc Hanbok, trang phục truyền thống của Hàn Quốc. Điều này cho thấy bức ảnh có thể được chụp ở Hàn Quốc và phản ánh văn hóa và truyền thống của gia đình này. Việc mặc Hanbok cũng có thể cho thấy sự tôn trọng đối với truyền thống và sự kiện quan trọng được chụp ảnh.
	Hai người đàn ông ngồi trên ghế bành đang làm gì? Họ có vẻ là ai?	Hai người đàn ông ngồi trên ghế bành có vẻ đang trong cuộc họp chính thức hoặc đàm phán. Họ có vẻ là những quan chức cấp cao của hai quốc gia (Cuba và Việt Nam), đang tham gia một cuộc gặp gỡ ngoại giao. Tư thế ngồi của họ thể hiện sự trang trọng và nghiêm túc.

Figure 3: Examples of input images, user questions, and generated answers from vRA-VLM.

of-domain scenarios.

- **RA-BLIP** (12): An extension of the BLIP-2 architecture, RA-BLIP incorporates a multimodal adaptive retrieval-augmented framework. It employs an adaptive selection knowledge generation strategy to filter and integrate relevant retrieved knowledge, improving performance in knowledge-intensive multimodal question-answering tasks.

These models exemplify the trend of combining VLMs with retrieval systems to enhance the models’ ability to generate accurate and contextually relevant responses. Although these approaches have shown promise in high-resource languages, their application in low-resource languages like Vietnamese remains underexplored. Our work aims to bridge this gap by developing a retrieval-augmented vision-language model tailored for Vietnamese, leveraging both visual and textual external knowledge to improve performance in multimodal tasks.

3. Proposal Method

In this section, we detail the architecture and training pipeline of our Vietnamese Retrieval-Augmented Vision-Language Model (vRA-VLM). Our approach consists of two main stages: (1) fine-tuning the model for detailed image description generation and (2) retrieval-augmented answer generation using external knowledge. The overall workflow of stage 1 is illustrated in Figure 1

Stage 1: Fine-tuning for Detailed Image Description Generation: We employ EVA-02 (4) as the vision encoder to extract robust visual representations from the input image. The encoder processes the image and outputs a sequence of visual feature embeddings \mathbf{v} , which serve as the foundation for subsequent fusion with language inputs.

A fixed instruction prompt in Vietnamese —"Miêu tả chi tiết bức ảnh này từ hành động, vật thể cho đến cảm giác hoặc bối cảnh. Bao gồm mọi yếu tố như màu sắc, ánh sáng, bố cục, các đối tượng và sự tương tác giữa chúng, cảm xúc hoặc thông điệp mà bức ảnh truyền tải. Hãy chú ý đến cả các yếu tố nhỏ mà có thể mang lại chiều sâu cho bức ảnh. Chỉ miêu tả những gì bạn nhận diện được, không được bịa đặt, ảo tưởng."— is tokenized and embedded using SeaLLMs, a Large Language Models for Southeast Asian Languages. This prompt ensures that the model consistently generates comprehensive and factual image descriptions

The prompt embeddings $\mathbf{p} \in R^{L_p \times d_p}$ and the visual patch embeddings $\mathbf{v} \in R^{L_v \times d_v}$ are fused via a cross-attention mechanism, which prompt tokens serve as queries, while the image patch embeddings are used as both keys and values. The cross-attention mechanism enables each prompt token to selectively attend to relevant visual regions, allowing model to integrate fine-grained visual information into the language presentation at each decoding step. Formally, the cross-attention is computed as:

$$\text{CrossAttn}(\mathbf{p}, \mathbf{v}) = \text{softmax} \left(\frac{\mathbf{p} \mathbf{W}_q (\mathbf{v} \mathbf{W}_k)^\top}{\sqrt{d_k}} \right) (\mathbf{v} \mathbf{W}_v)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learned projection matrices for the query, key, and value transformations, respectively. The fused embeddings are subsequently projected to match the input dimension of the language decoder

The fused representations are subsequently fed into the SeaLLMs decoder to autoregressively generate a detailed image description in Vietnamese. The model is fine-tuned using supervised learning with ground-truth captions, optimizing the cross-entropy loss between the predicted and

reference tokens:

$$\mathcal{L}_{\text{desc}} = - \sum_{t=1}^T \log P(y_t | y_{<t}, \mathbf{f})$$

where y_t is the target token at step t and \mathbf{f} is the fused representation from the cross-attention layer.

Stage 2: Retrieval-Augmented Answer Generation:

For each input sample, the image is first processed by the trained model from Stage 1 to generate a detailed image description. This ensures that all subsequent reasoning steps are grounded on a comprehensive, visual-centric context.

The generate description is then concatenated with the user question to form a composite textual query. This composite query is tokenized and embedding using model BAAI/bge-m3 (13), which is specifically designed for large-scale knowledge retrieval tasks. Unlike traditional embedding models limited by short input sequences, BGE-M3 supports a maximum input length of up to 8,192 tokens. This capability enables the model to encode lengthy knowledge base articles (such as Wikipedia passages) and composite queries that include both detailed image descriptions and user questions without truncation. This high token limit is particularly advantageous in our pipeline for two reasons:

- **Comprehensive Knowledge Encoding:** By embedding the entire content of long knowledge base entries, BGE-M3 preserves rich semantic and factual information, which improves the accuracy and coverage of retrieval results.
- **Rich Query Representation:** The ability to encode composite queries—including both the model-generated image description and the user question—ensures that all relevant context is considered during retrieval, resulting in more precise and contextually relevant matches.

The BGE-M3 model achieves strong performance on various retrieval benchmarks due to its multi-granular training (sentence, passage, and document level) and its robustness in multilingual and multi-domain scenarios. In our system, we leverage this capability to efficiently compute cosine similarity between the embedded composite query and the knowledge base, enabling accurate retrieval of supporting context for answer generation.

The final input to the language model is constructed by concatenating the retrieved context(s), the generated image description, and the user question in the following order: [description; question; context]. This sequence is tokenized into a single input for the language model.

The concatenated input is passed into the SeaLLMs language model, which is autoregressively generates the final answer. The model is trained using cross-entropy loss \mathcal{L}_{ans} with respect to the reference answers:

4. Dataset

5CD-AI/Viet-LAION-Gemini-VQA: We gratefully acknowledge the 5CD-AI (14) team for developing and publicly releasing the Viet-LAION-Gemini-VQA dataset, which serves as a critical resource enabling large-scale

Vietnamese vision-language research. Their effort in curating and maintaining this dataset has made this work possible.

This dataset is the large-scale open Vietnamese visual question answering resource to date, designed specifically to enable research in vision-language modeling for Vietnamese. It comprises approximately 843,529 samples in the train split, containing images from diverse domains, including journalism on social life, sports, tourism, e-commerce product images, clothing, vehicles, sketches, charts, games, technology, and more..., each associated with a rich set of annotations suitable for both image captioning and multi-turn vision-language conversation tasks. Each dataset entry includes the following fields:

- **id:** A unique string identifier for each sample.
- **image:** The visual input, stored as an image object.
- **description:** A detailed Vietnamese natural language description of the image, providing fine grained information about visible entities, attributes, actions, and scene context.
- **conversations:** A list of multi-turn vision-language conversations, where each conversation consists of a user question and a corresponding assistant answer, all in Vietnamese.

For Stage 1, we use the image-description pairs from the dataset. Each image is paired with its corresponding human-written description. This setup enables the model to learn to generate highly detailed and contextually accurate Vietnamese captions for diverse images, focusing on both object recognition and scene understanding. The descriptions are typically long-form and provide more than just object tags, often including attributes, relations, and background context. The examples of stage 1 image-description pairs are illustrated in Figure 2

For Stage 2, we utilize the image-description-conversation triples. Each sample includes not only the image and its description, but also a list of user-assistant conversation turns. In each conversation, a user poses a question about the image (sometimes requiring inference or reasoning beyond direct observation), and the assistant provides a corresponding answer. The examples of stage 2 image-question-conversation pairs are illustrated in Figure 3. For training, each conversation turn is treated as a separate training instance, where the model is conditioned on the image, its description, and the conversational history up to that point.

Vietnamese Wikipedia Knowledge Base: To support retrieval-augmented answer generation, we construct a knowledge base from the full dump of the Vietnamese Wikipedia. The knowledge base consists of millions of high-quality, human-curated articles spanning diverse domains such as history, science, geography, biographies, culture, and current events. Each article is segmented into passages or paragraphs, with each segment stored as an independent retrieval entry to improve granularity and retrieval precision.

All Wikipedia content is preprocessed to remove markup, tables, and non-informative text, retaining only plain Viet-

namese prose. Unicode normalization is applied to ensure consistent tokenization and encoding. The resulting corpus is indexed and embedded using the BAAI/bge-m3 model, which allows for efficient dense retrieval via cosine similarity between the composite query and the entire knowledge base.

The Vietnamese Wikipedia is selected due to its comprehensive, reliable coverage and its continuous updates by the Vietnamese-speaking community. Leveraging this knowledge base enables the system to ground answers in verified, up-to-date factual information, significantly reducing hallucination and improving factual consistency for Vietnamese multi-modal tasks.

5. Experimental Setup

All experiments are conducted on an NVIDIA H100 SXM (80GB VRAM) with Pytorch 2.7.0 and CUDA 12.6. Model development leverages the HuggingFace Transformers library and the timm vision backbones utilities.

Model Architecture: The vision encoder is instantiated from the pre-trained `eva02_small_patch14_336` (4) checkpoint, operating on 336×336 pixel RGB images. Images are normalized with mean $[0.481, 0.458, 0.408]$ and standard deviation $[0.269, 0.261, 0.276]$, following the ImageNet convention. The encoder outputs patch features from the penultimate layer (`select_layer: -2`), which serve as visual tokens for downstream processing. All weights of the vision encoder are kept frozen during training.

The language model backbone is `SeaLLMs-v3-1.5B` (15), with parameters initialized from the official checkpoint and load into BF16 precision for efficiency. The LLM is primarily kept frozen, only LoRA (16) adapters and the cross-attention fusion module are updated during fine-tuning. LoRA (Low-Rank Adaptation) adapters (16) are injected into all target modules of the language model with rank $r = 16$, $\alpha = 32$, and dropout 0.1.

The cross-attention module is configured with 8 attention heads, a feed-forward multiplier of 4, and dropout rate of 0.1.

Training Details: Training is performed for 10 epochs with a batch size of 2 and gradient accumulation steps of 16 (effective batch size 32). The AdamW optimizer is used with a learning rate of 1×10^{-5} , weight decay 0.01 and maximum gradient norm 1.0. All input and output sequences are truncated or padded to maximum length of 1536 tokens. The best model is selected based on the BLUE score on the validation set, with early stopping and `load_best_model_at_end` enabled.

All experiments are run with fixed random seeds where applicable. The source code, trained weights, and full training logs are made publicly available for reproducibility.

6. Discussion and Limitations

Due to technical limitations encountered with the HuggingFace Trainer API, we were unable to conduct evaluation on the held-out set. Specifically, integrating a custom `compute_metrics` function caused the Trainer to load the

Table I: Comparison of parameter quantity. Parameter quantities of other methods refer to (19), (21), (12).

Model	#Trainable Params	#Total Params
VLP+VinVL (17)	220M	220M
MuRAG (18)	527M	527M
SKURG (19)	447M	447M
ImplicitDecomp (20)	1310M	1.3B
RA-BLIP (T5-base) (12)	387M	1.4B
RA-BLIP (T5-large) (12)	902M	1.9B
RA-BLIP (FlanT5xl) (12)	109M	4.1B
RA-BLIP (FlanT5xxl) (12)	109M	12.1B
vRA-VLM	41M	8.1B

entire evaluation set into GPU memory during evaluation, resulting in CUDA out-of-memory (OOM) errors even on GPUs with very large memory capacity (up to 140GB). As a result, we could not report quantitative results at this stage.

This issue is a known limitation when using custom evaluation metrics with large-scale models and datasets, and may require redesigning the evaluation loop or implementing custom batch-wise metric computation outside the Trainer framework. By default, the Trainer’s metric computation is designed for convenience and speed, but is not scalable for large-scale evaluation workloads. Although the `eval_accumulation_steps` parameter can be set to periodically offload prediction tensors from GPU to CPU, this mechanism does not always resolve the issue, particularly when the custom metric function or collator is not optimized for streaming or mini-batch evaluation.

To overcome this limitation, it is necessary to redesign the evaluation pipeline so that metrics are computed in a streaming or batched manner—processing smaller chunks of predictions at a time and moving intermediate results to the CPU incrementally. This avoids aggregating the entire evaluation output in GPU memory. A fully custom evaluation loop, decoupled from the Trainer’s default `compute_metrics` interface, is the recommended solution for scaling evaluation to large datasets and complex multi-modal models.

Addressing this evaluation bottleneck is a priority for future work. Once resolved, we will be able to report comprehensive quantitative results and conduct deeper analyses of our model’s performance on downstream tasks.

Despite this limitation, we have established a robust training pipeline and demonstrated that the model can be fine-tuned efficiently on the available hardware. Once the evaluation bottleneck is addressed, we plan to report benchmark results and conduct in-depth analyses of model performance.

7. Future work

Given sufficient computational resources, we plan to extend our work by fine-tuning and evaluating the proposed architecture on a broader range of multimodal vision-language tasks in Vietnamese. Specifically, our future directions include:

- **OCR and Text Recognition:** We aim to adapt and benchmark the model on Vietnamese OCR and scene text understanding tasks using datasets such as **Viet-ViTextVQA-gemini-VQA**, **Viet-Vintext-gemini-VQA**, and **Viet-OCR-VQA**. This will assess the model’s ability to recognize and reason over textual content embedded within images.
- **Document Understanding** (Table II): We intend to evaluate the system on structured document understanding and reading comprehension tasks. Relevant datasets include **Viet-Doc-VQA** and **Viet-Doc-VQA-II** (general document VQA), **Viet-Geometry-VQA** and **Viet-ComputerScience-VQA** (subject-specific document reasoning), as well as **Viet-Sketches-VQA** (handwritten or diagram-based understanding).
- **Information Extraction:** We plan to investigate the model’s capability in information extraction from real-world documents, such as receipts and menus, by utilizing datasets like **Viet-Receipt-VQA** and **Viet-Menu-Gemini-VQA**.
- **Using world knowledge:** We plan to extend our evaluation to retrieval-augmented visual question answering tasks that require reasoning over external world knowledge beyond what is directly observable in the image. In particular, we aim to utilize datasets such as **AOKVQA**, which are specifically designed to benchmark models’ abilities to combine visual understanding with knowledge retrieved from large-scale text corpora or knowledge bases. Additionally, we will explore other retrieval-based multimodal QA datasets to further assess the model’s capacity for grounded, knowledge-intensive reasoning in open-domain and real-world scenarios.

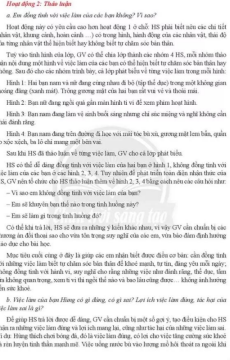
Through these extended experiments, we aim to systematically evaluate the generalization ability and versatility of the vRA-VLM architecture across diverse practical applications in the Vietnamese language ecosystem. Furthermore, insights gained from these tasks will guide subsequent improvements in model architecture and data preprocessing tailored for low-resource, domain-specific multimodal learning.

References

- [1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever "Learning Transferable Visual Models From Natural Language Supervision." *arXiv:2103.00020*, 2021 <https://arxiv.org/abs/2103.00020>
- [2] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." *arXiv:2201.12086*, 2022 <https://arxiv.org/abs/2201.12086>
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan "Flamingo: a Visual Language Model for Few-Shot Learning." *arXiv:2204.14198*, 2022 <https://arxiv.org/abs/2204.14198>
- [4] Yuxin Fang, Wen Wang, Xiaojie Lin, Jianmin Bao, Jian Zhang, Jingdong Shao, Jie Zhou, Wenguan Lu, Dongdong Yu, and Jian Dong. "EVA: Exploring the Limits of Masked Visual Representation Learning at Scale." *arXiv:2211.07636*, 2022. <https://arxiv.org/abs/2211.07636>
- [5] Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." *arXiv:2301.12597*, 2023 <https://arxiv.org/abs/2301.12597>
- [6] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee "Visual Instruction Tuning." *arXiv:2304.08485*, 2023 <https://arxiv.org/abs/2304.08485>
- [7] Davide Caffagni, Federico Cocchi, Nicholas Moratelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara "Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs." *arXiv:2404.15406*, 2024 <https://arxiv.org/abs/2404.15406>
- [8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, Douwe Kiela "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." *arXiv:2005.11401*, 2020 <https://arxiv.org/abs/2005.11401>
- [9] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, Ming-Wei Chang "REALM: Retrieval-Augmented Language Model Pre-Training." *arXiv:2002.08909*, 2020 <https://arxiv.org/abs/2002.08909>
- [10] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, Laurent Sifre "Improving language models by retrieving from trillions of tokens." *arXiv:2112.04426*, 2021 <https://arxiv.org/abs/2112.04426>
- [11] Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Ming-Yu Liu, Yuke Zhu, Mohammad Shoeybi, Bryan Catanzaro, Chaowei Xiao,

- Anima Anandkumar "Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning." *arXiv:2302.04858*, 2023 <https://arxiv.org/abs/2302.04858>
- [12] Muhe Ding, Yang Ma, Pengda Qin, Jianlong Wu, Yuhong Li, Liqiang Nie "RA-BLIP: Multi-modal Adaptive Retrieval-Augmented Bootstrapping Language-Image Pre-training." *arXiv:2410.14154*, 2024 <https://arxiv.org/abs/2410.14154>
- [13] Xiaonan Li, Zhenyu Hou, Yuning Hong, Yucheng Fu, Jinghui Liu, Junlong Li, Yiyang Wen, Yaobo Liang, Xiaodong Liu, Yasheng Wang, Qun Liu, Haizhou Li. "BAAI General Embedding (BGE): An Embedding Model Bridging Text, Code, and Table." *arXiv:2402.03216*, 2024 <https://arxiv.org/pdf/2402.03216>
- [14] Khang Doan, Bao Huynh Gia, Dung Hoang Tien, Dinh Pham Thuc, Quan Nguyen Tran Minh, Bang Quoc Vo, Suong Nhat Hoang, Huynh Pham Nhat, Tien Le Minh Fifth Civil Defender - 5CD <https://huggingface.co/5CD-AI>
- [15] Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, Lidong Bing "SeaLLMs - Language Models for Southeast Asian Languages." *arXiv:2312.00738*, 2023 <https://arxiv.org/abs/2312.00738>
- [16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen "LoRA: Low-Rank Adaptation of Large Language Models." *arXiv:2106.09685*, 2021 <https://arxiv.org/abs/2106.09685>
- [17] Y. Chang, G. Cao, M. Narang, J. Gao, H. Suzuki, and Y. Bisk "WebQA: Multihop and Multimodal QA." *arXiv:2109.00590*, 2021 <https://arxiv.org/abs/2109.00590>
- [18] Wenhui Chen, Hexiang Hu, Xi Chen, Pat Verga, William W. Cohen "MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text." *arXiv:2210.02928*, 2022 <https://arxiv.org/abs/2210.02928>
- [19] Qian Yang, Qian Chen, Wen Wang, Baotian Hu, Min Zhang "Enhancing Multi-modal and Multihop Question Answering via Structured Knowledge and Unified Retrieval-Generation." *arXiv:2212.08632*, 2022 <https://arxiv.org/abs/2212.08632>
- [20] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, Jonathan Berant "MultiModalQA: Complex Question Answering over Text, Tables and Images." *arXiv:2104.06039*, 2021 <https://arxiv.org/abs/2104.06039>
- [21] Bowen Yu, Cheng Fu, Haiyang Yu, Fei Huang, Yongbin Li "Unified Language Representation for Question Answering over Text, Tables, and Images."

Table II: An example of image, description, and conversation pairs of document understanding datasets.

Image	Description	Conversations
	<p>Ảnh là một trang sách giáo khoa, có tiêu đề "Hoạt động 2: Thảo luận". Nội dung chính của trang sách là thảo luận về việc làm của các bạn trong hình ảnh. Ảnh có các hình ảnh minh họa cho các tình huống được thảo luận. Nội dung văn bản được chia thành hai phần chính, được đánh dấu bằng chữ cái a và b:</p> <p>**Phần a:** Bắt đầu với câu hỏi "Em đồng tình với việc làm của các bạn không? Vì sao?" Sau đó, phần này cung cấp một số tình huống được minh họa bằng hình ảnh, ví dụ: Hai bạn nam và nữ đang cùng nhau đi bộ (tập thể dục) trong một khoảng thời gian, Bạn nữ đang ngồi quá gần màn hình tivi để xem phim hoạt hình, Bạn nam đang làm vệ sinh buổi sáng nhưng chỉ súc miệng và nghiêng người, Bạn nam đang trên đường đi học với mái tóc bù xù, gương mặt lem bẫn, quần áo xộc xệch, ba lô chỉ mang một bên vai. Sau khi học sinh thảo luận về tình việc làm, giáo viên cho cả lớp phát biểu. Phần này cũng đưa ra các câu hỏi để học sinh thảo luận thêm về các hình ảnh, chẳng hạn: "Vì sao em không đồng tình với việc làm của bạn?", "Em sẽ khuyên bạn thế nào trong tình huống này?", "Có thể làm gì trong tình huống đó?" Cuối cùng, phần này nêu mục tiêu bài học là giúp học sinh nhận biết được điều cơ bản là cần đồng tình với những việc làm biết tự chăm sóc bản thân để khỏe mạnh, tự tin, dạng yêu mỗi ngày, không đồng tình với hành vi, suy nghĩ cho rằng những việc như danh rằng, thể dục, tắm rửa không quan trọng, xem ti vi thì ngồi thế nào và bao lâu cũng được...không ảnh hưởng đến sức khỏe.</p> <p>**Phần b:** Bắt đầu với câu hỏi "Việc làm của bạn Hùng có gì đúng, có gì sai? Lợi ích việc làm đúng, tác hại của việc làm sai là gì?" Phần này cung cấp câu trả lời cho giáo viên để hướng dẫn học sinh, ví dụ: Để giúp học sinh trả lời được dễ dàng, giáo viên cần chuẩn bị một số gợi ý, tạo điều kiện cho học sinh nhận ra những việc làm đúng và lợi ích mang lại, cũng như tác hại của những việc làm không đúng, đó là việc làm đúng và lợi ích mang lại, cũng tác hại của những việc làm sai. Ví dụ: Hùng thích chơi bóng đá, đó là việc làm đúng, có lợi cho tăng cường sức khỏe và rèn luyện tinh thần mạnh mẽ. Việc uống nước bổ sung lượng nước bị thoát ra ngoài khi...</p>	<p>Q: Theo ảnh, học sinh cần thảo luận về vấn đề gì?</p> <p>A: Học sinh cần thảo luận về việc làm của các bạn trong hình ảnh, cụ thể là đồng tình hay không đồng tình với việc làm của các bạn trong mỗi hình ảnh, và lý do tại sao.</p> <p>Q: Trong phần a, có bao nhiêu tình huống được minh họa bằng hình ảnh?</p> <p>A: Có 4 tình huống được minh họa bằng hình ảnh trong phần a.</p> <p>Q: Mục tiêu của bài học này là gì?</p> <p>A: Mục tiêu của bài học là giúp học sinh nhận biết được điều cơ bản là cần đồng tình với những việc làm biết tự chăm sóc bản thân để khỏe mạnh, tự tin, dạng yêu mỗi ngày, không đồng tình với hành vi, suy nghĩ cho rằng những việc như danh rằng, thể dục, tắm rửa không quan trọng, xem ti vi thì ngồi thế nào và bao lâu cũng được...không ảnh hưởng đến sức khỏe.</p> <p>Q: Theo ảnh, việc làm sai của bạn Hùng có thể là gì?</p> <p>A: Ảnh không cung cấp thông tin cụ thể về việc làm sai của bạn Hùng, nhưng phần b đề cập đến việc làm sai là việc không đồng tình với những việc làm đúng và lợi ích mang lại, cũng như tác hại của những việc làm sai. Ví dụ: không đồng tình với những việc như danh rằng, thể dục, tắm rửa không quan trọng, xem ti vi thì ngồi thế nào và bao lâu cũng được...không ảnh hưởng đến sức khỏe.</p>

<p>Hình ảnh thể hiện một hình tròn với tâm O và các điểm A, B, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, X, E1, F1 được đánh dấu trên hình tròn và các đường thẳng nối các điểm này. Bên cạnh hình vẽ là nội dung văn bản trình bày chứng minh cho hai khẳng định:</p> <p>1. Các tứ giác BFHD, CEHD, BFEC nội tiếp. Chứng minh: Do AD, BE, CF là các đường cao của tam giác ABC nên HDB = BFD = 90°. => HDB + BFD = 180°, suy ra tứ giác BFHD nội tiếp (tổng hai góc đối nhau bằng 180°). Tương tự ta cũng có tứ giác CEHD nội tiếp. Ta có: BFC = BEC nên BFEC là tứ giác nội tiếp (Hai đỉnh liên tiếp F, E cùng nhìn cạnh BC một góc bằng nhau).</p> <p>2. Các đường thẳng AD, BE, CF chứa các đường phân giác của góc EDF; DEF; EFD, từ đó suy ra trực tâm H là tâm đường tròn nội tiếp tam giác DEF. Chứng minh: Vì BFHD nội tiếp nên: FBH = FDH (cùng chắn FH) (1), CEHD nội tiếp nên HDE = HCE (cùng chắn EH) (2), tứ giác BFEC nội tiếp nên FBE = FCE (cùng chắn EF) (3).</p>	<p>Hình ảnh thể hiện một hình tròn với tâm O và các điểm A, B, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, X, E1, F1 được đánh dấu trên hình tròn và các đường thẳng nối các điểm này. Bên cạnh hình vẽ là nội dung văn bản trình bày chứng minh cho hai khẳng định:</p> <p>1. Các tứ giác BFHD, CEHD, BFEC nội tiếp. Chứng minh: Do AD, BE, CF là các đường cao của tam giác ABC nên HDB = BFD = 90°. => HDB + BFD = 180°, suy ra tứ giác BFHD nội tiếp (tổng hai góc đối nhau bằng 180°). Tương tự ta cũng có tứ giác CEHD nội tiếp. Ta có: BFC = BEC nên BFEC là tứ giác nội tiếp (Hai đỉnh liên tiếp F, E cùng nhìn cạnh BC một góc bằng nhau).</p> <p>2. Các đường thẳng AD, BE, CF chứa các đường phân giác của góc EDF; DEF; EFD, từ đó suy ra trực tâm H là tâm đường tròn nội tiếp tam giác DEF. Chứng minh: Vì BFHD nội tiếp nên: FBH = FDH (cùng chắn FH) (1), CEHD nội tiếp nên HDE = HCE (cùng chắn EH) (2), tứ giác BFEC nội tiếp nên FBE = FCE (cùng chắn EF) (3).</p>	<p>Q: Trong hình vẽ, điểm nào là tâm của đường tròn? A: Điểm O là tâm của đường tròn. Q: Tại sao tứ giác BFHD nội tiếp? A: Tứ giác BFHD nội tiếp vì tổng hai góc đối nhau HDB và BFD bằng 180°. Cụ thể, HDB = BFD = 90° vì AD, BE, CF là các đường cao của tam giác ABC. Q: Tại sao BFEC là tứ giác nội tiếp? A: FEC là tứ giác nội tiếp vì hai đỉnh liên tiếp F, E cùng nhìn cạnh BC một góc bằng nhau, tức là góc BFC bằng góc BEC. Q: Tại sao H là tâm đường tròn nội tiếp tam giác DEF? A: H là tâm đường tròn nội tiếp tam giác DEF vì các đường thẳng AD, BE, CF chứa các đường phân giác của góc EDF; DEF; EFD. Điều này được chứng minh bằng cách sử dụng các cặp góc bằng nhau trong các tứ giác nội tiếp: FBH = FDH, HDE = HCE, FBE = FCE.</p>
<p>Hình ảnh thể hiện một hình tròn với tâm O và các điểm A, B, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, X, E1, F1 được đánh dấu trên hình tròn và các đường thẳng nối các điểm này. Bên cạnh hình vẽ là nội dung văn bản trình bày chứng minh cho hai khẳng định:</p> <p>1. Các tứ giác BFHD, CEHD, BFEC nội tiếp. Chứng minh: Do AD, BE, CF là các đường cao của tam giác ABC nên HDB = BFD = 90°. => HDB + BFD = 180°, suy ra tứ giác BFHD nội tiếp (tổng hai góc đối nhau bằng 180°). Tương tự ta cũng có tứ giác CEHD nội tiếp. Ta có: BFC = BEC nên BFEC là tứ giác nội tiếp (Hai đỉnh liên tiếp F, E cùng nhìn cạnh BC một góc bằng nhau).</p> <p>2. Các đường thẳng AD, BE, CF chứa các đường phân giác của góc EDF; DEF; EFD, từ đó suy ra trực tâm H là tâm đường tròn nội tiếp tam giác DEF. Chứng minh: Vì BFHD nội tiếp nên: FBH = FDH (cùng chắn FH) (1), CEHD nội tiếp nên HDE = HCE (cùng chắn EH) (2), tứ giác BFEC nội tiếp nên FBE = FCE (cùng chắn EF) (3).</p>	<p>Bức ảnh bao gồm 30 hình vẽ được sắp xếp thành 6 hàng, mỗi hàng có 5 hình vẽ. Các hình vẽ được đánh số từ 1 đến 30, theo thứ tự từ trái sang phải và từ trên xuống dưới. Các hình vẽ đa dạng, bao gồm các đồ vật, động vật, con người, và các biểu tượng trừu tượng. Phong cách vẽ đơn giản, sử dụng nét vẽ mảnh, tạo cảm giác hỗn nhiên và ngây thơ.</p>	<p>Q: Hình vẽ số 3 là gì và bạn có suy đoán gì về nó? A: Hình vẽ số 3 là một tấm bảng với dòng chữ "Happy Birthday" được viết bằng tiếng Nhật. Có thể suy đoán rằng đây là một hình vẽ về một tấm bảng chúc mừng sinh nhật, tạo nên một bầu không khí vui tươi và ấm áp. Q: Hình vẽ số 8 và 9 mô tả những gì? Có mối liên hệ nào giữa chúng? A: Hình vẽ số 8 là một con ong, hình vẽ số 9 là một con gấu. Mỗi liên hệ giữa chúng có thể là cả hai đều là động vật, một là côn trùng, một là động vật có vú, và có thể biểu thị sự tương phản giữa thế giới côn trùng và thế giới động vật. Q: Hình vẽ số 13 và 14 thể hiện những gì? Ý nghĩa của chúng là gì? A: Hình vẽ số 13 là một chiếc lược, hình vẽ số 14 là một thanh kiếm. Cả hai đều là những vật dụng quen thuộc, nhưng lại đại diện cho hai khía cạnh đối lập: một là sự gọn gàng, chăm chút bản thân, còn một là sự mạnh mẽ, uy quyền. Q: Hình vẽ số 13 và 14 thể hiện những gì? Ý nghĩa của chúng là gì? A: Hình vẽ số 13 là một chiếc lược, hình vẽ số 14 là một thanh kiếm. Cả hai đều là những vật dụng quen thuộc, nhưng lại đại diện cho hai khía cạnh đối lập: một là sự gọn gàng, chăm chút bản thân, còn một là sự mạnh mẽ, uy quyền.</p>