ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

NGÀNH KHOA HỌC MÁY TÍNH

Đề tài

TRỢ LÝ ẢO GIỚI THIỆU DI SẢN VĂN HÓA CÁCH MẠNG TÂY NAM BỘ DỰA TRÊN HÌNH ẢNH VỚI PHOBERT KẾT HỢP VIT5

IMAGE-BASED VIRTUAL ASSISTANT FOR INTRODUCING THE REVOLUTIONARY CULTURAL HERITAGE OF THE SOUTHWEST REGION USING PHOBERT AND VIT5

Sinh viên: Lê Thị Ngọc Ngân

Mã số: B2106801

Khoá: 47

Giảng viên hướng dẫn 1: TS. Trần Nguyễn Minh Thư

Giảng viên hướng dẫn 2: Ths. Huỳnh Gia Khương

Cần Thơ, 12/2025

ĐẠI HỌC CẦN THƠ

TRƯỜNG CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

NGÀNH KHOA HỌC MÁY TÍNH

Đề tài

TRỢ LÝ ẢO GIỚI THIỆU DI SẢN VĂN HÓA CÁCH MẠNG TÂY NAM BỘ DỰA TRÊN HÌNH ẢNH VỚI PHOBERT KẾT HỢP VIT5

IMAGE-BASED VIRTUAL ASSISTANT FOR INTRODUCING THE REVOLUTIONARY CULTURAL HERITAGE OF THE SOUTHWEST REGION USING PHOBERT AND VIT5

Giảng viên hướng dẫn

TS. Trần Nguyễn Minh Thư

Ths. Huỳnh Gia Khương

Sinh viên thực hiện

Họ và tên: Lê Thị Ngọc Ngân

Mã số: B2106801

Khoá: 47

Cần Thơ, 12/2025

LỜI CẢM ƠN

Để hoàn thành luận văn này, bên cạnh sự nỗ lực của bản thân, tôi đã nhận được sự quan tâm, hỗ trợ và giúp đỡ quý báu từ nhiều tập thể và cá nhân.

Trước hết, tôi xin bày tỏ lòng biết ơn sâu sắc đến Cô TS. Trần Nguyễn Minh Thư và Thầy Ths. Huỳnh Gia Khương đã tận tình hướng dẫn, chỉ bảo và đồng hành cùng tôi trong suốt quá trình nghiên cứu và thực hiện đề tài.

Tôi xin chân trọng cảm ơn quý Thầy, Cô Trường Công nghệ Thông tin và Truyền thông - Trường Đại học Cần Thơ, những người đã tận tâm truyền đạt tri thức và nhiệt huyết, tạo nền tảng vững chắc cho tôi trong suốt quá trình học tập tại trường.

Tôi cũng xin gửi lời cảm ơn sâu sắc đến gia đình thân yêu, những người đã luôn làm điểm tựa vững chắc, động viên và ủng hộ tôi vượt qua những khó khăn trong học tập và cuộc sống.

Bên cạnh đó, tôi xin chân thành cảm ơn các bạn bè đã đồng hành, chia sẻ và khích lệ tôi, góp phần giúp tôi hoàn thành tốt luận văn này.

Trong quá trình thực hiện, luận văn khó tránh khỏi những thiếu sót và hạn chế, kính mong nhận được sự chỉ dẫn, góp ý của quý Thầy, Cô để luận văn được hoàn thiện hơn.

Xin trân trọng cảm ơn!

Cần Thơ, ngày……tháng……năm 2025

Tác giả

Lê Thị Ngọc Ngân

MỤC LỤC

DANH MỤC HÌNH ẢNH

DANH MỤC BẢNG BIỂU

DANH MỤC TỪ VIẾT TẮT

STT	Từ viết tắc (Abbreviation)	Từ viết tắc đầy đủ (Origin word)
1	VQA	Visual Question Answering
2	VIT	Vision Transformer
3	PEFT	Parameter-Efficient Fine-Tuning
4	RAG	Retrieval-Augmented Generation
5	NLP	Natural Language Processing
6	CNN	Convolutional Neural Network
7	RNN	Recurrent Neural Network
8	LSTM	Long Short-Term Memory
9	PE	Positional Encoding
10	BPE	Byte Pair Encoding
11	MLP	Multi-Layer Perceptron
12	VLMs	Vision-Language Models
13	LLM	Large Language Model
14	EM	Exact Match
15	BP	Brevity Penalty
16	LCS	Longest Common Subsequence

ABSTRACT

This thesis originates from the need to provide cultural information through a question–answering format, aiming to support the introduction of revolutionary cultural heritage sites in the Southwest region of Vietnam. The system enables users to interact directly in Vietnamese by posing questions about images of historical monuments and receiving contextually appropriate answers.

The study investigates two main architectural approaches: (i) a Modular framework integrating Vision Transformer (ViT), PhoBERT, and ViT5; and (ii) a Unified architecture built upon the Qwen2-VL model. Both models were trained on a custom Vietnamese VQA dataset consisting of 1,226 images from 81 heritage sites across five provinces in the Southwest region, along with 5,631 question–answer pairs. To optimize computational resources, the thesis adopts Parameter-Efficient Fine-Tuning (PEFT) through the LoRA technique, reducing fine-tuning costs to only 0.1–0.3% of the total model parameters.

Experimental results show that the Modular architecture achieves a BERTScore-F1 of 0.81 and a CIDEr score of 0.79, while the Unified architecture reaches a BERTScore-F1 of 0.80 and CIDEr of 0.76, demonstrating the models' strong semantic alignment with ground-truth answers. The system is integrated into an intuitive web interface supporting image upload, question submission, and automatic answer generation, contributing to heritage preservation, education, and digital promotion.

Keywords: Visual Question Answering, cultural heritage, Southwest Vietnam, Vision Transformer, PhoBERT, ViT5, Qwen2-VL, LoRA, virtual assistant.

TÓM TẮT

Xuất phát từ nhu cầu cung cấp thông tin văn hóa qua hình thức hỏi–đáp, hệ thống được xây dựng nhằm hỗ trợ giới thiệu các di sản văn hóa cách mạng tại khu vực Tây Nam Bộ. Hệ thống cho phép người dùng tương tác trực tiếp bằng tiếng Việt thông qua việc đặt câu hỏi về hình ảnh các di tích lịch sử và nhận lại câu trả lời phù hợp với ngữ cảnh.

Nghiên cứu đề xuất và đánh giá hai hướng tiếp cận: (i) kiến trúc Modular kết hợp Vision Transformer (ViT), PhoBERT và ViT5; và (ii) kiến trúc Unified dựa trên mô hình Qwen2-VL. Cả hai mô hình được huấn luyện trên bộ dữ liệu VQA tiếng Việt được xây dựng riêng, gồm 1.226 hình ảnh của 81 di tích thuộc năm tỉnh Tây Nam Bộ và 5.631 cặp câu hỏi – câu trả lời. Để tối ưu tài nguyên tính toán, luận văn áp dụng phương pháp Parameter-Efficient Fine-Tuning (PEFT) với kỹ thuật LoRA, giúp giảm chi phí tinh chỉnh xuống còn 0,1–0,3% tổng số tham số mô hình.

Kết quả thực nghiệm cho thấy mô hình Modular đạt BERTScore-F1 0,81 và CIDEr 0,79, trong khi mô hình Unified đạt BERTScore-F1 0,80 và CIDEr 0,76, chứng minh khả năng sinh câu trả lời có mức tương đồng ngữ nghĩa cao với tham chiếu. Hệ thống được tích hợp vào giao diện web trực quan, hỗ trợ tải ảnh, đặt câu hỏi và nhận câu trả lời tự động, qua đó góp phần vào công tác bảo tồn, giáo dục và quảng bá di sản trong thời kỳ chuyển đổi số.

Từ khóa: Visual Question Answering, di sản văn hóa cách mạng, Tây Nam Bộ, Vision Transformer, PhoBERT, ViT5, Qwen2-VL, LoRA, trợ lý ảo.

I. PHẦN GIỚI THIỆU

1. Đặt vấn đề

Trong bối cảnh chuyển đổi số đang diễn ra mạnh mẽ trên toàn cầu, việc ứng dụng các công nghệ số vào bảo tồn và phát huy giá trị di sản văn hóa trở thành một xu hướng tất yếu. Các giải pháp số hóa không chỉ góp phần lưu giữ và lan tỏa bản sắc văn hóa dân tộc, mà còn tạo điều kiện thuận lợi để thế hệ trẻ tiếp cận thông tin lịch sử một cách trực quan và hiệu quả hơn.

Tại Việt Nam, đặc biệt ở khu vực Tây Nam Bộ, hệ thống di sản văn hóa cách mạng có giá trị lịch sử sâu sắc, tiêu biểu như Khu di tích Xẻo Quýt (Đồng Tháp), Căn cứ Trung ương Cục miền Nam (Tây Ninh) hay Căn cứ Rừng U Minh (Cà Mau). Tuy nhiên, công tác giới thiệu và quảng bá các di sản này vẫn còn hạn chế. Thông tin chủ yếu được cung cấp thông qua bảng giới thiệu, video tư liệu hoặc bài viết tĩnh trên website – những hình thức mang tính một chiều, ít tương tác và chưa tận dụng tốt các tiến bộ công nghệ nhằm nâng cao trải nghiệm người dùng.

Sự phát triển nhanh chóng của trí tuệ nhân tạo, đặc biệt trong các lĩnh vực xử lý ngôn ngữ tự nhiên (NLP), thị giác máy tính (Computer Vision) và học sâu (Deep Learning), đã mở ra nhiều hướng tiếp cận mới cho các hệ thống tương tác thông minh. Trong đó, Visual Question Answering (VQA) – công nghệ cho phép hệ thống trả lời câu hỏi dựa trên nội dung hình ảnh – đã chứng minh tiềm năng ứng dụng trong giáo dục, bảo tàng và truyền thông văn hóa, nhờ khả năng kết hợp giữa yếu tố trực quan và ngôn ngữ để tạo ra trải nghiệm tương tác phong phú.

Mặc dù vậy, khảo sát các nghiên cứu hiện nay cho thấy các hệ thống VQA chủ yếu tập trung vào bài toán tổng quát, sử dụng các bộ dữ liệu quy mô lớn như COCO-QA và VQA v2, hoặc được ứng dụng trong thương mại và giáo dục phổ thông. Tại Việt Nam, đặc biệt trong lĩnh vực bảo tồn và giới thiệu di sản cách mạng, chưa có nghiên cứu hoặc sản phẩm công bố nào ứng dụng mô hình VQA để kết hợp hình ảnh di tích với ngôn ngữ tự nhiên nhằm cung cấp thông tin lịch sử theo hình thức tương tác.

Xuất phát từ khoảng trống này, luận văn lựa chọn đề tài ***"Trợ lý ảo giới thiệu di sản văn hóa cách mạng Tây Nam Bộ dựa trên hình ảnh"***, với mục tiêu xây dựng một hệ thống có khả năng kết hợp xử lý ảnh và xử lý ngôn ngữ tự nhiên để cung cấp thông tin về di sản một cách trực quan, sinh động và thân thiện với người dùng. Về mặt thực tiễn, hệ thống có thể triển khai tại bảo tàng, khu di tích lịch sử, cơ sở giáo dục hoặc các nền tảng số phục vụ quảng bá văn hóa địa phương, góp phần nâng cao hiệu quả truyền thông di sản và hỗ trợ giáo dục lịch sử theo định hướng đổi mới.

2. Những nghiên cứu liên quan

Một số nghiên cứu trước đây đã đề cập đến bài toán Visual Question Answering (VQA) trong nhiều lĩnh vực khác nhau. Mặc dù chưa có công trình nào tập trung trực tiếp vào bài toán ***"Trợ lý ảo giới thiệu di sản văn hóa cách mạng Tây Nam Bộ dựa trên hình ảnh"***, các nghiên cứu này vẫn cung cấp cơ sở lý thuyết và phương pháp quan trọng cho đề tài.

Trong số các công trình có liên quan, Bongini và cộng sự [1] là nhóm tác giả đầu tiên triển khai VQA trong bối cảnh di sản văn hóa. Nghiên cứu tập trung xây dựng hệ thống trả lời câu hỏi dựa trên hình ảnh tác phẩm nghệ thuật, kết hợp giữa đặc trưng thị giác trích xuất bằng mạng neuron tích chập (CNN) và mô hình xử lý ngôn ngữ tự nhiên (NLP) để hiểu câu hỏi của người dùng. Kết quả cho thấy việc tích hợp hai nguồn thông tin – hình ảnh và văn bản – có thể tăng độ chính xác của hệ thống trong bối cảnh đặc thù của lĩnh vực văn hóa – nghệ thuật.

Becattini và cộng sự [2] mở rộng hướng tiếp cận này bằng việc xây dựng VISCOUNTH – bộ dữ liệu VQA đa ngôn ngữ quy mô lớn dành riêng cho lĩnh vực di sản văn hóa. VISCOUNTH bao gồm khoảng 500.000 hình ảnh và 6,5 triệu cặp câu hỏi – câu trả lời bằng tiếng Anh và tiếng Ý, được tạo ra bằng phương pháp bán tự động dựa trên đồ thị tri thức ArCo kết hợp mẫu câu hỏi từ chuyên gia và người dùng. Các câu trả lời được tạo tự động thông qua truy vấn vào đồ thị tri thức. Kết quả thử nghiệm trên các mô hình VQA hiện đại đạt trung bình 0,61 F1 và 0,51 Exact Match, chứng minh giá trị của bộ dữ liệu trong nghiên cứu VQA chuyên biệt theo lĩnh vực. Tuy nhiên, các mô hình vẫn gặp khó khăn đáng kể với các câu hỏi yêu cầu suy luận ngữ nghĩa ở mức cao hoặc liên kết thông tin ngữ cảnh phức tạp.

Bên cạnh các nghiên cứu theo hướng lĩnh vực, cộng đồng nghiên cứu VQA nói chung đã phát triển nhiều mô hình đa phương thức hiệu quả. ViLT [3], một kiến trúc Vision-and-Language Transformer tối giản, loại bỏ hoàn toàn tầng CNN và đạt 76,13% trên NLVR2 và 71,26% trên VQAv2. BLIP [4], với chiến lược Bootstrapping Language-Image Pre-training và 14 triệu hình ảnh huấn luyện, đạt 77,54% trên VQA test-dev và 77,62% trên test-std. ViLBERT [5] và LXMERT [6], các kiến trúc song mã (dual-stream) sử dụng hai encoder cho hình ảnh và ngôn ngữ, đạt lần lượt 70,92% và 72,5% trên VQA v2, cùng hiệu suất cao trên các bộ đánh giá đa nhiệm khác như GQA và NLVR2. Những mô hình này chứng minh sự tiến bộ đáng kể của các kỹ thuật học sâu đa phương thức trong việc kết hợp thông tin hình ảnh và văn bản.

Tại Việt Nam, một số nghiên cứu đã bước đầu tiếp cận VQA trong ngữ cảnh tiếng Việt. Mô hình BARTPhoBEiT [7] đạt độ chính xác 68,58% trên bộ dữ liệu ViVQA thông qua việc kết hợp mô hình sinh ngôn ngữ BARTPho với mô hình thị giác BEiT. ViCAN [8], tiếp cận theo hướng kết hợp thông tin đa tầng (multi-level fusion), đạt 68,76% trên ViVQA. Nghiên cứu "Advancing Vietnamese Visual Question Answering

with Transformer and Convolutional Integration" [9] đạt kết quả 71,04% trên cùng bộ dữ liệu, cho thấy việc kết hợp Transformer và CNN giúp tăng cường khả năng mã hóa thị giác và cải thiện độ chính xác.

Nhìn chung, các nghiên cứu quốc tế đã chứng minh tiềm năng của VQA trong lĩnh vực văn hóa, song chủ yếu tập trung vào tác phẩm nghệ thuật hoặc dữ liệu tổng quát, trong khi các nghiên cứu tại Việt Nam mới chỉ thử nghiệm trên bộ dữ liệu ViVQA – vốn không phản ánh đặc trưng của hình ảnh di sản văn hóa cách mạng. Bên cạnh đó, chưa có bộ dữ liệu hay mô hình nào được thiết kế riêng cho bối cảnh di sản Việt Nam, đặc biệt là khu vực Tây Nam Bộ. Khoảng trống này là cơ sở quan trọng để đề tài xây dựng một hệ thống VQA chuyên biệt, phù hợp với nội dung di sản và đặc thù ngôn ngữ – ngữ cảnh tiếng Việt.

3. Mục tiêu đề tài

Mục tiêu chính của đề tài là thiết kế và triển khai một hệ thống trợ lý ảo ứng dụng mô hình Visual Question Answering (VQA), nhằm hỗ trợ người dùng truy xuất, tìm hiểu và tương tác trực tiếp với các di sản văn hóa cách mạng tại khu vực Tây Nam Bộ thông qua việc đặt câu hỏi dựa trên hình ảnh. Hệ thống được kỳ vọng nâng cao tính trực quan trong trải nghiệm người dùng, cải thiện khả năng tiếp cận thông tin, đồng thời đóng góp vào công tác bảo tồn, giáo dục và phát huy giá trị di sản trong bối cảnh chuyển đổi số.

Bên cạnh mục tiêu ứng dụng, đề tài hướng tới việc khảo sát, phân tích và đánh giá các phương pháp tiếp cận VQA hiện có, bao gồm các kiến trúc Modular và Unified, từ đó xác định hướng giải pháp tối ưu cho việc triển khai mô hình trợ lý ảo trên dữ liệu di sản văn hóa, đảm bảo hiệu quả và khả năng mở rộng trong các bối cảnh tương tác thực tiễn.

4. Đối tượng và phạm vi nghiên cứu

a. Đối tượng nghiên cứu

Đối tượng nghiên cứu của đề tài bao gồm bài toán VQA và các hướng tiếp cận liên quan trong lĩnh vực trí tuệ nhân tạo, đặc biệt tập trung vào các mô hình học sâu đa phương thức (multi-modal deep learning) kết hợp xử lý hình ảnh và ngôn ngữ tự nhiên. Đồng thời, đề tài nghiên cứu nguồn tư liệu về di sản văn hóa cách mạng khu vực Tây Nam Bộ, được sử dụng làm tập dữ liệu chính cho quá trình xây dựng, huấn luyện và đánh giá mô hình.

b. Phạm vi nghiên cứu

Phạm vi nghiên cứu được giới hạn trong việc ứng dụng mô hình VQA nhằm nhận diện và trả lời các câu hỏi liên quan đến di sản văn hóa cách mạng tại khu vực Tây Nam

Bộ. Dữ liệu nghiên cứu bao gồm hình ảnh và văn bản mô tả bằng tiếng Việt về các di sản tiêu biểu thuộc các tỉnh Cần Thơ, Hậu Giang, Sóc Trăng, An Giang và Kiên Giang.

Trong phạm vi này, đề tài tập trung vào việc xây dựng và đánh giá một mô hình trợ lý ảo có khả năng hỗ trợ người dùng truy xuất thông tin và tương tác trực quan với nội dung di sản thông qua hình ảnh

5. Phương pháp nghiên cứu

Phương pháp nghiên cứu của đề tài bao gồm các bước: (i) thu thập, phân tích và tổng hợp các tài liệu, công trình khoa học liên quan đến bài toán VQA và các mô hình học sâu đa phương thức; (ii) xây dựng tập dữ liệu chuyên biệt về di sản văn hóa cách mạng khu vực Tây Nam Bộ, bao gồm hình ảnh và văn bản mô tả; (iii) thiết kế, triển khai mô hình trợ lý ảo dựa trên các kiến trúc VQA phù hợp; và (iv) kiểm thử, đánh giá hiệu quả hệ thống dựa trên các tiêu chí định lượng và định tính, đảm bảo tính chính xác, khả năng tương tác và phù hợp với mục tiêu ứng dụng thực tiễn. Phương pháp này kết hợp giữa nghiên cứu lý thuyết và thực nghiệm, nhằm tối ưu hóa hiệu suất mô hình và độ tin cậy của hệ thống trong môi trường thực tế.

6. Bố cục quyển báo cáo

Quyển báo cáo được tổ chức thành ba phần chính:

- **I. Phần giới thiệu:** Trình bày lý do chọn đề tài, tổng quan các nghiên cứu liên quan, xác định mục tiêu nghiên cứu, đối tượng – phạm vi nghiên cứu và phương pháp được sử dụng trong quá trình thực hiện luận văn.
- **II. Phần nội dung:** Gồm ba chương chính:
  - **Chương 1. Mô tả bài toán:** Trình bày chi tiết bài toán nghiên cứu và các vấn đề giải pháp có liên quan.
  - **Chương 2. Thiết kế và cài đặt giải pháp:** Mô tả quy trình xây dựng tập dữ liệu, thiết kế mô hình và triển khai hệ thống trợ lí ảo dựa trên công nghệ VQA.
  - **Chương 3. Kiểm thử và đánh giá:** Trình bày phương pháp đánh giá, các phép đo được sử dụng, kết quả thực nghiệm và phân tích hiệu quả của mô hình và hệ thống.
- **III. Phần kết luận:** Tóm tắt những kết quả đạt được, nêu rõ các đóng góp chính của đề tài, chỉ ra những hạn chế còn tồn tại và đề xuất các hướng nghiên cứu – phát triển trong tương lai.

II. PHẦN NỘI DUNG

CHƯƠNG 1. MÔ TẢ BÀI TOÁN

1.1. Mô tả chi tiết bài toán

Bài toán nghiên cứu tập trung vào việc thiết kế và triển khai một hệ thống trợ lý ảo dựa trên mô hình Visual Question Answering (VQA) nhằm hỗ trợ việc giới thiệu và khám phá các di sản văn hóa cách mạng tại khu vực Tây Nam Bộ. Hệ thống được xây dựng để tiếp nhận đồng thời hai loại dữ liệu đầu vào: (i) hình ảnh các đối tượng di sản văn hóa cách mạng và (ii) câu hỏi tiếng Việt do người dùng đặt ra. Dựa trên các đầu vào này, hệ thống thực hiện trích xuất đặc trưng hình ảnh, biểu diễn ngôn ngữ tự nhiên, tích hợp thông tin đa phương thức và sinh câu trả lời dạng văn bản ngắn gọn, chính xác, phù hợp với ngữ cảnh hình ảnh.

Mục tiêu chính của nghiên cứu là nâng cao trải nghiệm tương tác người dùng và khắc phục những hạn chế của các phương thức truyền tải thông tin truyền thống vốn mang tính tĩnh, thiếu trực quan và không hỗ trợ tương tác hai chiều. Thông qua việc tích hợp các kỹ thuật trí tuệ nhân tạo, thị giác máy tính và xử lý ngôn ngữ tự nhiên, hệ thống được kỳ vọng trở thành công cụ hiệu quả trong việc truyền thông, giáo dục và nâng cao nhận thức cộng đồng về giá trị lịch sử – văn hóa của vùng Tây Nam Bộ.

Về mặt kỹ thuật, hệ thống được yêu cầu thực hiện quy trình xử lý đa phương thức với ba thành phần chính:

- **Trích xuất đặc trưng hình ảnh:** Phân tích nội dung trực quan của hình ảnh di sản, đảm bảo hiệu quả và ổn định trong nhiều điều kiện chụp khác nhau.
- **Biểu diễn ngôn ngữ tự nhiên:** Xử lý câu hỏi tiếng Việt, bao gồm các cách diễn đạt đa dạng, ngữ cảnh phức tạp và các thuật ngữ chuyên ngành lịch sử – văn hóa.
- **Tích hợp thông tin đa phương thức và suy luận:** Kết hợp dữ liệu hình ảnh và ngôn ngữ để sinh câu trả lời chính xác, có ý nghĩa và phù hợp với bối cảnh câu hỏi.

Bài toán nghiên cứu đặt ra một số thách thức kỹ thuật, bao gồm:

- Quy mô dữ liệu hạn chế và tính đa dạng chưa cao, đặc biệt đối với hình ảnh di sản lịch sử.
- Độ phức tạp của xử lý ngôn ngữ tiếng Việt, nhất là với các cụm từ chuyên môn hoặc tên riêng lịch sử.

- Yêu cầu cao về độ chính xác, tốc độ phản hồi và khả năng hiểu đúng ngữ cảnh.
- Khả năng mở rộng hệ thống khi bổ sung dữ liệu mới mà không làm giảm hiệu năng.

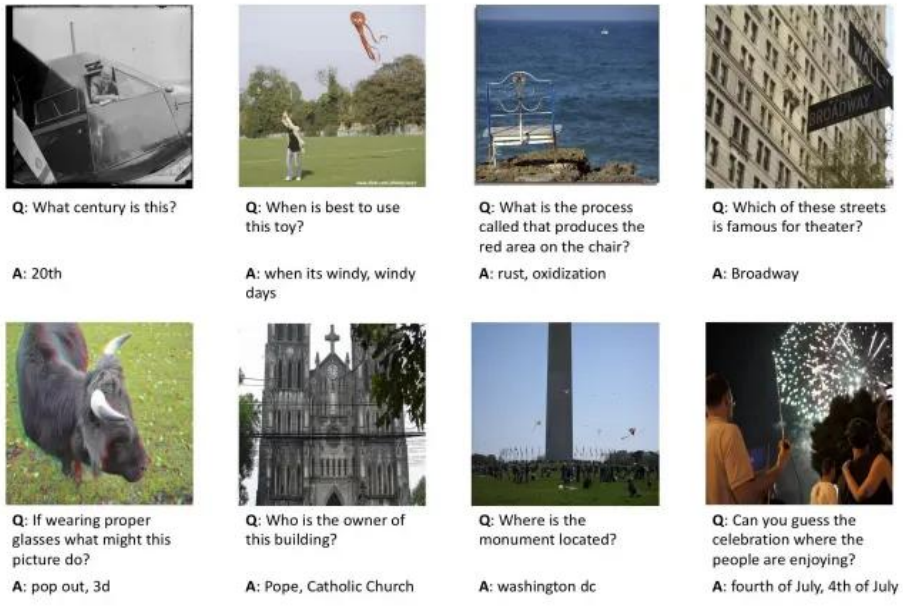
Hướng tiếp cận được đề xuất là ứng dụng các mô hình học sâu đa phương thức, kết hợp bộ mã hóa hình ảnh và bộ mã hóa ngôn ngữ trong một mô hình VQA thống nhất để sinh câu trả lời chính xác và phù hợp. Bộ dữ liệu nghiên cứu bao gồm ảnh di sản, mô tả văn bản và tập câu hỏi – đáp án chuẩn hóa, đảm bảo chất lượng cho quá trình huấn luyện và đánh giá mô hình.

Về phạm vi ứng dụng, hệ thống không chỉ phục vụ hoạt động giới thiệu và giáo dục tại bảo tàng, khu di tích hay nền tảng số mà còn có khả năng mở rộng sang các loại hình di sản văn hóa khác, đóng góp vào công tác bảo tồn, truyền thông và giáo dục trong bối cảnh chuyển đổi số hiện nay.

1.2. Vấn đề giải pháp liên quan đến bài toán

1.2.1. Visual Question Answering (VQA)

Visual Question Answering (VQA) là một hướng nghiên cứu quan trọng trong lĩnh vực trí tuệ nhân tạo đa phương thức, trong đó hệ thống được yêu cầu sinh câu trả lời A dựa trên thông tin thu nhận từ ảnh I và câu hỏi văn bản Q (*Hình 1*). Bài toán thường được mô hình hóa dưới dạng phân phối xác suất có điều kiện $P(A \mid I, Q)$, cho phép mô hình học cách suy luận và tối ưu hóa lựa chọn câu trả lời từ hai nguồn dữ liệu đầu vào. VQA là một tác vụ mang tính tương tác cao, đòi hỏi sự kết hợp chặt chẽ giữa nhận thức thị giác, hiểu ngôn ngữ tự nhiên và lập luận đa phương thức, nhằm đảm bảo câu trả lời vừa chính xác vừa phù hợp với ngữ cảnh hình ảnh.



Hình 1: Minh họa bài toán Visual Question Answering (VQA)

Bài toán VQA vẫn là một thách thức do tính đa phương thức, tồn tại cả ở miền thị giác và miền ngôn ngữ. Ở miền thị giác, mô hình phải thực hiện các nhiệm vụ nhận diện đối tượng, xác định thuộc tính, quan hệ không gian và bối cảnh tổng thể; bất kỳ sai lệch nào trong biểu diễn hình ảnh đều có thể lan truyền và làm suy giảm chất lượng suy luận. Ở miền ngôn ngữ, hệ thống cần phân tích cú pháp, giải đoán mục đích câu hỏi và xác định loại thông tin mà người dùng yêu cầu. Đồng thời, các dạng câu hỏi trong VQA rất đa dạng, từ câu hỏi đếm, so sánh đến các câu hỏi mô tả mở, khiến việc trích xuất ý nghĩa trở nên phức tạp. Thách thức cốt lõi của VQA nằm ở khả năng lập luận đa phương thức, tức là quá trình liên kết, căn chỉnh và tích hợp thông tin giữa đặc trưng hình ảnh và ngôn ngữ một cách nhất quán. Đối với các mô hình sinh câu trả lời mở (open-ended generation), yêu cầu tạo ra văn bản tự nhiên, ngắn gọn và đảm bảo tính mạch lạc ngữ nghĩa càng làm tăng độ phức tạp của bài toán.

Hiện nay, hai hướng tiếp cận chủ đạo trong nghiên cứu VQA là mô hình dạng module (Modular Approach) và mô hình hợp nhất (Unified/End-to-End Approach). Hướng tiếp cận dạng module phân rã bài toán thành các giai đoạn tương đối độc lập, bao gồm trích xuất đặc trưng thị giác, phân tích ngôn ngữ, hợp nhất thông tin và sinh câu trả lời. Ưu điểm của cách tiếp cận này là khả năng diễn giải cao, dễ tùy chỉnh từng thành phần và phù hợp với các bài toán có dữ liệu hạn chế. Tuy nhiên, do các giai đoạn có quan hệ phụ thuộc, sai lệch ở một bước có thể lan truyền và ảnh hưởng tiêu cực đến toàn bộ hệ thống (error propagation).

Ngược lại, hướng tiếp cận hợp nhất tối ưu toàn bộ quy trình trong một kiến trúc duy nhất. Các mô hình này học biểu diễn chung và tối ưu đồng thời trên cả miền thị giác và ngôn ngữ, nhờ đó thường đạt hiệu quả suy luận cao hơn. Tuy nhiên, chúng mang tính hộp đen (black-box), khó phân tích cơ chế hoạt động và đòi hỏi lượng dữ liệu lớn để đảm bảo tính ổn định và khả năng tổng quát hóa.

Trong nghiên cứu này, hướng tiếp cận dạng module được lựa chọn làm mô hình chính, dựa trên các lý do sau:

- Nguồn dữ liệu VQA tiếng Việt, đặc biệt trong lĩnh vực di sản văn hóa, còn hạn chế, khiến các mô hình end-to-end khó đạt hiệu quả tối ưu.
- Modular Approach cho phép tận dụng các tài nguyên xử lý tiếng Việt đã được nghiên cứu trước, bao gồm các mô hình phân tích ngôn ngữ, hệ thống nhận dạng văn bản tiếng Việt và bộ phân loại câu hỏi.
- Cấu trúc dạng module giúp kiểm soát tính đúng đắn của từng bước xử lý, đồng thời tạo điều kiện thuận lợi cho việc tích hợp tri thức miền (domain knowledge), đặc biệt quan trọng khi xử lý các đối tượng đặc thù trong lĩnh vực di sản văn hóa.

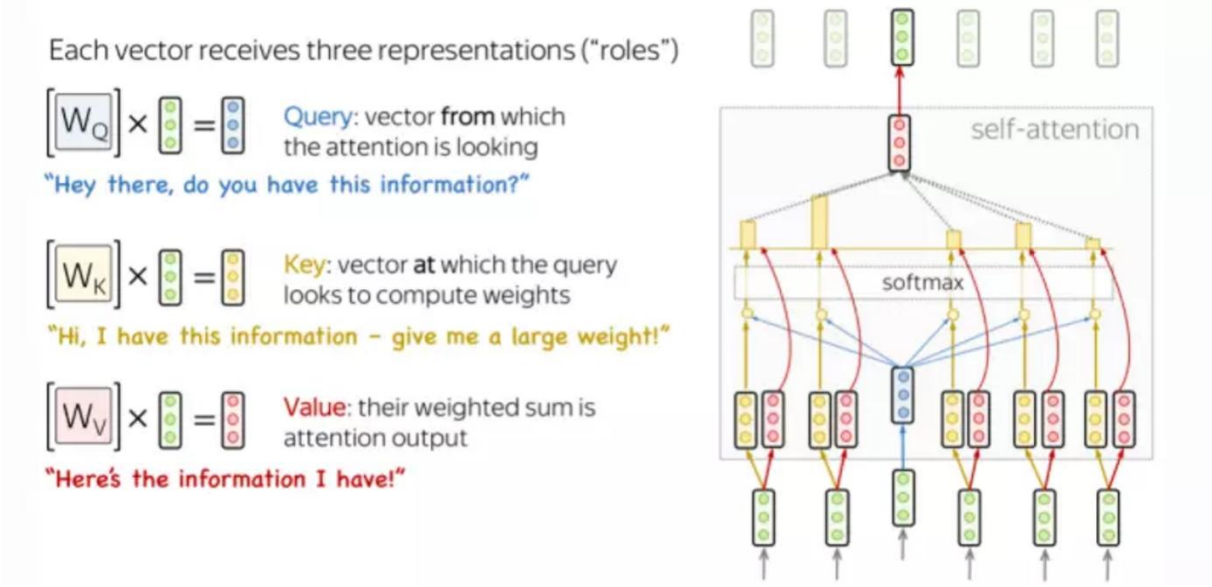
Nhờ những ưu điểm trên, hướng tiếp cận dạng module mang lại tính linh hoạt và độ tin cậy cao cho hệ thống VQA trong điều kiện dữ liệu và ngữ cảnh ứng dụng cụ thể của đề tài.

1.2.2. Cơ chế Attention trong Transformer

Các mô hình xử lý chuỗi truyền thống như RNN và LSTM [10] gặp nhiều hạn chế trong khả năng học biểu diễn. Thứ nhất, hiện tượng vanishing gradient làm suy giảm năng lực ghi nhớ thông tin dài hạn, khiến mô hình khó nắm bắt các ngữ cảnh xa. Thứ hai, cơ chế xử lý tuần tự (sequential processing) không cho phép song song hóa, dẫn đến thời gian huấn luyện kéo dài và hiệu suất xử lý thấp. Bên cạnh đó, việc mô hình hóa các phụ thuộc xa (long-range dependencies) cũng gặp khó khăn, do trạng thái phải lan truyền qua nhiều bước thời gian. Những hạn chế này là động lực ra đời của Transformer – một kiến trúc hoàn toàn dựa trên cơ chế Attention, vừa cải thiện khả năng biểu diễn vừa tối ưu hóa tốc độ huấn luyện.

1.2.2.1. Cơ chế Self-Attention

Self-Attention [11] là thành phần cốt lõi của Transformer, cho phép mô hình học mối quan hệ giữa các phần tử trong cùng một chuỗi mà không bị ràng buộc bởi khoảng cách vị trí. Mỗi phần tử đầu vào được ánh xạ thành ba vector: $Query$ (Q), Key (K) và $Value$ (V) thông qua các phép biến đổi tuyến tính. Mức độ liên quan giữa các phần tử được xác định bằng cách tính tích vô hướng giữa Q và K; kết quả sau đó được chuẩn hóa bằng hàm softmax và dùng để trọng số hóa Value, tạo ra biểu diễn mới (*Hình 2*):



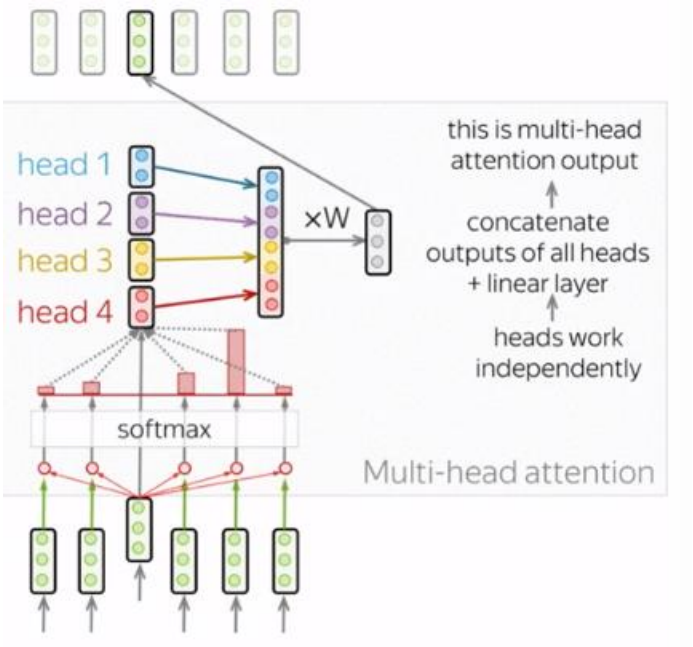
Hình 2: : Cơ chế Self-Attention trong Transformer

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Trong đó d_k là số chiều của vector Key, được dùng để chuẩn hóa nhằm tránh giá trị tích vô hướng quá lớn khi không gian biểu diễn tăng. Kết quả của Self-Attention phản ánh mức độ đóng góp của từng phần tử trong chuỗi đối với phần tử đang xét, giúp mô hình định trọng số các thông tin quan trọng mà không phụ thuộc vào cơ chế lan truyền tuần tự như trong RNN/LSTM.

1.2.2.2. Multi-Head Attention

Để khai thác thông tin từ nhiều không gian biểu diễn khác nhau, Transformer áp dụng Multi-Head Attention [11]. Mỗi "head" thực hiện một phép Self-Attention độc lập với bộ tham số riêng, tạo ra các biểu diễn khác nhau của cùng một chuỗi đầu vào. Các kết quả từ các head sau đó được nối lại và biến đổi qua ma trận chiếu tuyến tính W^O (*Hình 3*):



Hình 3: Cơ chế Multi-Head Attention trong Transformer

$$MultiHead(Q,K,V) = Concat(\text{head}_1, \dots, \text{head}_n)W^O$$

Cơ chế này cho phép mô hình đồng thời học nhiều dạng quan hệ khác nhau, bao gồm quan hệ cú pháp, ngữ nghĩa hoặc các quy luật hình thái trong ngôn ngữ. Nghiên cứu thực nghiệm cho thấy mỗi head thường tập trung vào các cấu trúc hoặc phụ thuộc ngữ cảnh khác nhau, từ đó tăng cường khả năng biểu diễn và hiệu quả xử lý các tác vụ đa tầng thông tin.

1.2.2.3. Positional Encoding

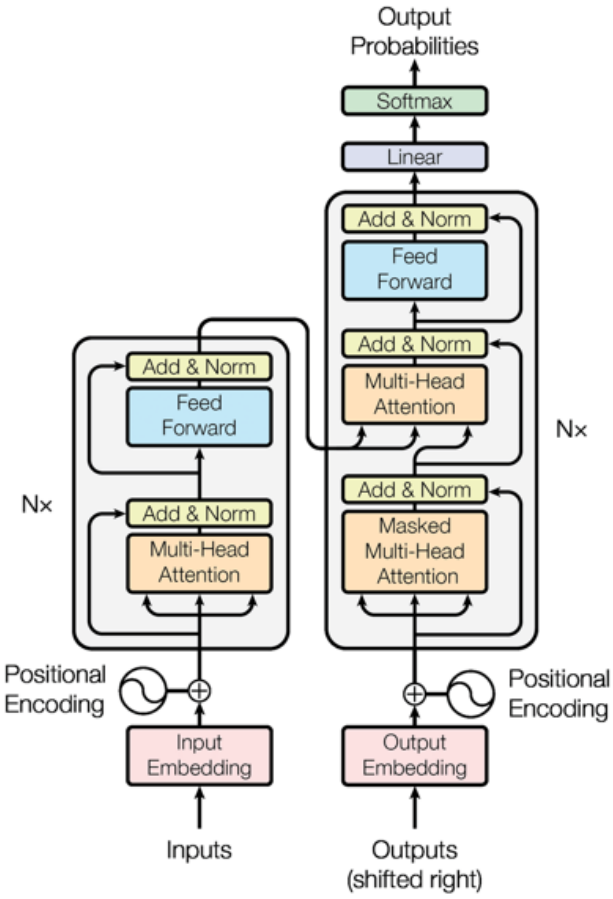
Một hạn chế cố hữu của Self-Attention là tính order-invariant, tức mô hình không tự động nhận biết thứ tự các phần tử. Để bổ sung thông tin vị trí, Transformer sử dụng Positional Encoding (PE) [11], dựa trên các hàm sin/cos với các tần số khác nhau:

$$PE(pos, 2i) = sin\left(\frac{pos}{1000^{\frac{2i}{d_{model}}}}\right),$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{1000^{2i/d_{model}}}\right)$$

Trong đó pos là vị trí phần tử trong chuỗi và i là chỉ số chiều của vector. Thiết kế này cho phép mô hình học mối quan hệ tương đối giữa các vị trí, đồng thời duy trì khả năng khái quát hóa cho các chuỗi dài hơn so với dữ liệu huấn luyện.

1.2.2.4. Kiến trúc Encoder và Decoder của Transformer



Hình 4: Kiến trúc Encoder-Decoder của mô hình Transformer

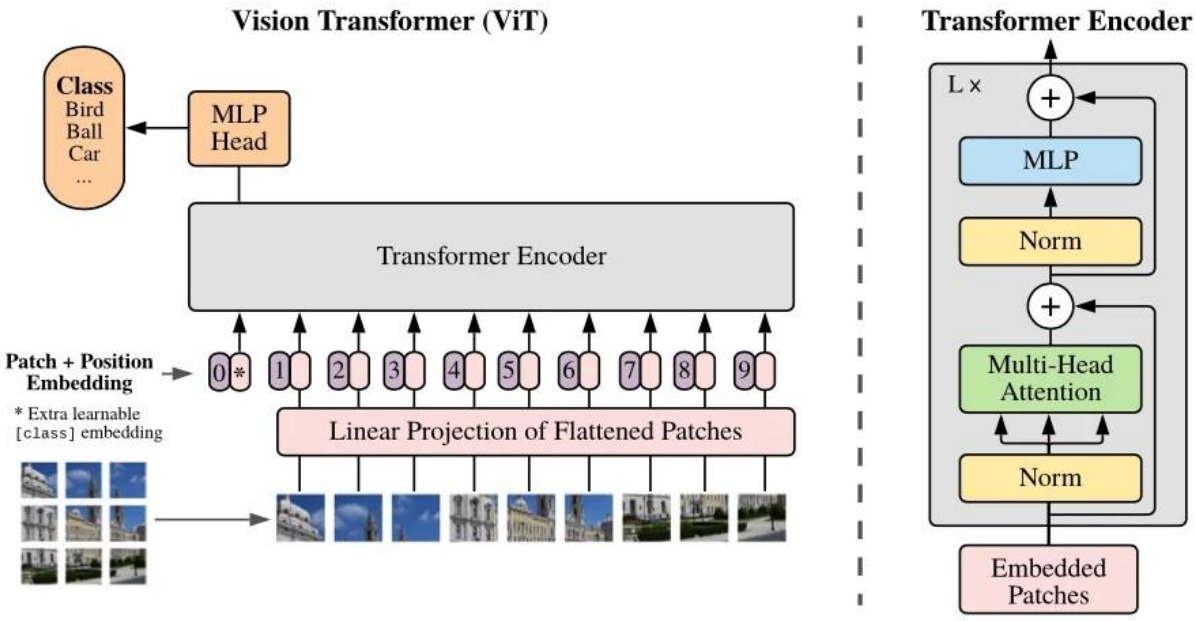
(Nguồn: Vaswani et al., 2017 [11])

Transformer sử dụng kiến trúc Encoder–Decoder [11], trong đó mỗi thành phần gồm nhiều lớp lặp lại chứa các mô-đun Multi-Head Attention và Feed-Forward Networks (*Hình 4*). Encoder trích xuất và tổng hợp thông tin ngữ cảnh từ chuỗi đầu vào, trong khi Decoder kết hợp thông tin đã được mã hóa với cơ chế masked self-attention để sinh chuỗi đầu ra một cách có kiểm soát. Nhờ kết hợp Attention đa hướng và khả năng xử lý song song, Transformer đạt hiệu quả vượt trội trong nhiều tác vụ NLP và thị giác máy tính, bao gồm cả Visual Question Answering (VQA).

1.2.3. Vision Transformer cho Visual Understanding

Vision Transformer (ViT) [12] được đề xuất như một hướng tiếp cận mới trong xử lý ảnh, trong đó mô hình áp dụng kiến trúc Transformer – vốn thành công trong xử lý ngôn ngữ tự nhiên – trực tiếp lên dữ liệu hình ảnh. Cách tiếp cận này giúp mô hình học được mối quan hệ toàn cục giữa các vùng ảnh, khắc phục các hạn chế cố hữu của mạng tích chập truyền thống (CNN).

Trước hết, CNN [13] được xây dựng dựa trên hai đặc tính cảm ứng (inductive bias) quan trọng: locality và translation invariance. Điều này cho phép CNN học tốt các đặc trưng cục bộ và duy trì tính ổn định khi đối tượng dịch chuyển trong ảnh. Tuy nhiên, chính đặc tính này cũng khiến CNN gặp khó khăn trong việc mô hình hóa các quan hệ dài hạn hoặc cấu trúc tổng thể của ảnh, do receptive field tăng chậm theo số lượng tầng tích chập. Việc yêu cầu nhiều tầng để thu thập thông tin toàn cục vừa làm tăng chi phí tính toán, vừa khiến mô hình dễ mất thông tin chi tiết.



Hình 5: Kiến trúc Vision Transformer (ViT)

(Dosovitskiy et al., 2020 [12])

Để khắc phục vấn đề này, ViT đề xuất cơ chế Image Patching, trong đó ảnh đầu vào được chia thành các ô vuông (patch) không chồng lấn có kích thước cố định. Ví dụ, một ảnh 224×224 được chia thành 196 patch kích thước 16×16. Mỗi patch sau đó được làm phẳng (flatten) thành một vector có kích thước $P^2 \times C$ và tạo thành một chuỗi N phần tử, với N là số lượng patch. Bằng cách này, hình ảnh được chuyển đổi thành dạng chuỗi tương tự như chuỗi token văn bản trong NLP, giúp mô hình Transformer có thể xử lý trực tiếp.

Tiếp theo, mỗi patch được ánh xạ sang không gian biểu diễn có chiều cố định thông qua phép chiếu tuyến tính, tạo thành patch embeddings. Mô hình cũng bổ sung một class token vào đầu chuỗi để tổng hợp thông tin đặc trưng phục vụ các nhiệm vụ phân loại. Bên cạnh đó, để mô hình có thể nhận biết vị trí tương đối giữa các patch, ViT sử dụng position embeddings dưới dạng tham số học được, cho phép mô hình giữ lại cấu trúc không gian của ảnh.

Kiến trúc chính của ViT bao gồm một chuỗi Transformer Encoder gồm L tầng. Mỗi tầng bao gồm cơ chế self-attention đa đầu (multi-head self-attention) và các khối feed-forward phi tuyến. Thông qua đó, mô hình có khả năng học các quan hệ phụ thuộc toàn cục giữa mọi patch trong ảnh, thay vì chỉ giới hạn trong vùng lân cận như CNN. Output cuối cùng của ViT là tập đặc trưng hình ảnh $F_v \in R^{N \times d}$, trong đó d là kích thước không gian embedding.

Nhờ khả năng mô hình hóa quan hệ toàn cục và tính linh hoạt trong xử lý độ phân giải, ViT đặc biệt phù hợp cho các bài toán yêu cầu reasoning ở mức cao như Visual Question Answering (VQA). Cơ chế self-attention cho phép mô hình chú ý đồng thời đến nhiều vùng quan trọng trong ảnh, thay vì bị ràng buộc vào cấu trúc cục bộ của CNN. Bên cạnh đó, ViT thường đạt hiệu năng cao khi được tiền huấn luyện trên các tập dữ liệu quy mô lớn, giúp tăng chất lượng biểu diễn thị giác trong các tác vụ downstream. Những đặc điểm này góp phần nâng cao khả năng hiểu nội dung hình ảnh và hỗ trợ quá trình suy luận đa phương thức trong hệ thống VQA.

1.2.4. Pre-trained Language Models cho tiếng Việt

Trong bối cảnh tiếng Việt thuộc nhóm ngôn ngữ có tài nguyên hạn chế (low-resource), việc xây dựng các mô hình học sâu từ đầu gặp nhiều khó khăn do thiếu dữ liệu huấn luyện và chi phí tính toán cao. Do đó, việc sử dụng các mô hình ngôn ngữ tiền huấn luyện (pre-trained language models) đã trở thành một hướng tiếp cận hiệu quả, giúp cải thiện chất lượng biểu diễn ngôn ngữ và giảm yêu cầu về dữ liệu gán nhãn trong các tác vụ hạ nguồn.

Đối với bài toán Visual Question Answering (VQA) tiếng Việt, các mô hình tiền huấn luyện đóng vai trò quan trọng trong việc mã hóa câu hỏi và sinh câu trả lời tự nhiên, chính xác. Trong nghiên cứu này, hai mô hình được lựa chọn là PhoBERT cho bước mã hóa câu hỏi và ViT5 cho bước sinh câu trả lời, nhờ khả năng biểu diễn ngữ nghĩa tốt và đã được chứng minh hiệu quả trong các tác vụ xử lý tiếng Việt.

1.2.4.1. PhoBERT – Mô hình mã hóa câu hỏi

PhoBERT [14] là mô hình pre-trained tiêu biểu dành riêng cho tiếng Việt, phát triển dựa trên kiến trúc RoBERTa-base. Mô hình bao gồm 12 lớp Transformer, với kích thước vector ẩn 768 chiều và 12 đầu attention, được huấn luyện trên khoảng 20GB dữ liệu văn bản tiếng Việt tổng hợp từ nhiều nguồn như báo chí, Wikipedia và các kho ngữ

liệu lớn. Quá trình huấn luyện trên tập dữ liệu đa dạng này cho phép PhoBERT đạt chất lượng biểu diễn ngôn ngữ vượt trội so với các mô hình đa ngôn ngữ như mBERT hay XLM-R trong hầu hết các tác vụ NLP tiếng Việt.

Một đặc thù khi xử lý tiếng Việt là vấn đề phân tách từ, do nhiều từ đơn và từ ghép được viết cách nhau bằng dấu cách, khiến các tokenizer mặc định khó nhận biết ranh giới từ. PhoBERT giải quyết vấn đề này bằng cách kết hợp với các công cụ như RDRSegmenter, VnCoreNLP, và PyVi để thực hiện word segmentation trước khi áp dụng Byte Pair Encoding (BPE). Quá trình này chuẩn hóa các cụm từ đa từ như "học sinh", "quốc gia", "nhân dân" thành các token dạng "học_sinh", "quốc_gia", "nhân_dân", từ đó cải thiện chất lượng biểu diễn ngữ nghĩa.

Trong bối cảnh VQA, PhoBERT được sử dụng để mã hóa câu hỏi đầu vào. Với một câu hỏi gồm T token sau khi tách từ và mã hóa, PhoBERT sinh ra ma trận embedding $Q \in R^{T\times 768}$. Token đặc biệt $[CLS]$, được thiết kế để nắm bắt ngữ nghĩa tổng thể của câu, được sử dụng làm vector đại diện cho câu hỏi trong các bước hợp nhất thông tin giữa ảnh và văn bản. Nhờ khả năng biểu diễn ngữ nghĩa mạnh, PhoBERT cung cấp đầu vào ổn định và giàu thông tin cho các bước xử lý tiếp theo trong hệ thống VQA.

1.2.4.2. ViT5 – Mô hình sinh câu trả lời

ViT5 [15] là phiên bản tiếng Việt của kiến trúc T5, một mô hình encoder–decoder thiết kế theo hướng text-to-text, trong đó mọi bài toán NLP được ánh xạ về dạng chuỗi sang chuỗi (sequence-to-sequence). ViT5 được huấn luyện trên tập dữ liệu lớn tiếng Việt, với khoảng 32.000 token được xây dựng theo kỹ thuật SentencePiece BPE, đảm bảo bao phủ các biến thể từ vựng trong ngôn ngữ tự nhiên.

ViT5 được phát hành với nhiều kích thước khác nhau. Hai phiên bản phổ biến là base (~220 triệu tham số) và large (~750 triệu tham số). Cả hai phiên bản giữ nguyên kiến trúc T5 gốc, bao gồm cơ chế self-attention trong encoder, cross-attention trong decoder, và mục tiêu tiền huấn luyện dạng phục hồi chuỗi bị che khuất (span corruption). Cấu trúc này cho phép mô hình học được các quan hệ phụ thuộc dài hạn và sinh văn bản tự nhiên phù hợp với ngữ cảnh.

Trong hệ thống VQA tiếng Việt, ViT5 đảm nhiệm vai trò sinh câu trả lời dựa trên thông tin nhận được từ encoder văn bản và đặc trưng hình ảnh. Quá trình sinh đầu ra được thực hiện theo cơ chế auto-regressive, tức mỗi token được mô hình dự đoán dựa trên toàn bộ token đã sinh trước đó. Các phương pháp giải mã như beam search hoặc top-k sampling có thể được áp dụng để cải thiện độ trôi chảy và tự nhiên của câu trả lời.

So với các mô hình đa ngôn ngữ, ViT5 thể hiện ưu thế rõ rệt trong việc sinh ngôn ngữ tiếng Việt tự nhiên, mạch lạc và nhất quán với ngữ cảnh. Đồng thời, ViT5 cho hiệu quả tốt khi fine-tuning với lượng dữ liệu vừa phải, phù hợp với điều kiện thực tế của

nhiều bài toán tiếng Việt, bao gồm cả VQA. Nhờ những đặc điểm này, ViT5 là lựa chọn phù hợp để xây dựng thành phần sinh câu trả lời trong hệ thống đề xuất.

1.2.5. Kỹ thuật Fusion đa phương thức (Cross-Modal Fusion)

Trong các hệ thống Visual Question Answering (VQA), việc kết hợp thông tin từ hai miền dữ liệu dị biệt – ảnh và văn bản – đóng vai trò trung tâm trong việc hình thành câu trả lời chính xác. Hai nguồn dữ liệu này vốn được biểu diễn trong những không gian đặc trưng không tương đồng: hình ảnh được mô tả dưới dạng tín hiệu thị giác (pixel-level hoặc patch-level features), trong khi văn bản được biểu diễn trong không gian ngữ nghĩa rời rạc của từ và câu. Sự khác biệt này tạo ra một "khoảng cách ngữ nghĩa" (semantic gap), khiến mô hình gặp khó khăn trong việc thiết lập mối tương quan giữa các đối tượng trong ảnh và những thành phần ngôn ngữ trong câu hỏi. Do đó, các kỹ thuật kết hợp đa phương thức (multimodal fusion) xuất hiện nhằm thu hẹp khoảng cách này và cung cấp cơ chế học tương tác chặt chẽ giữa hai modality.

Trong các phương pháp truyền thống, hai chiến lược fusion phổ biến nhất là early fusion và late fusion. Early fusion thực hiện ghép nối trực tiếp các vector đặc trưng hình ảnh và văn bản, sau đó đưa vào một mạng nơ-ron nhiều lớp (MLP) để dự đoán. Cách tiếp cận này đơn giản và dễ triển khai, nhưng không cho phép mô hình học được các mối quan hệ phức tạp giữa vùng ảnh và thành phần của câu hỏi. Ngược lại, late fusion xử lý từng modality độc lập bằng các module chuyên biệt (ví dụ: CNN/Vision Transformer cho ảnh và Transformer cho văn bản), rồi chỉ kết hợp kết quả ở tầng cuối. Mặc dù giảm được nhiễu không cần thiết giữa hai miền dữ liệu, phương pháp này lại hạn chế khả năng mô hình hóa các tương tác chi tiết ở mức từ–vật thể (word–object alignment), vốn rất quan trọng trong VQA.

Để khắc phục những hạn chế trên, các mô hình hiện đại chuyển sang sử dụng cơ chế cross-modal attention – một dạng mở rộng của cơ chế attention trong Transformer. Thay vì kết hợp hai modality theo cách tuyến tính, cross-attention cho phép một modality "tập trung" (attend) vào modality còn lại thông qua cơ chế truy vấn–khóa–giá trị (Query–Key–Value). Khi văn bản đóng vai trò chủ đạo (text-guided attention), các token thuộc câu hỏi được ánh xạ thành Query, trong khi các đặc trưng thị giác được ánh xạ thành Key và Value. Quá trình attention giúp mô hình xác định những vùng ảnh có liên quan đến nội dung câu hỏi, ví dụ như từ "màu" thường kích hoạt những vùng chứa đối tượng có màu sắc nổi bật trong ảnh. Ngược lại, trong trường hợp cần hội tụ thông tin từ cả hai phía, mô hình có thể triển khai cơ chế attention hai chiều (bidirectional attention), trong đó cả ảnh và câu hỏi thay phiên nhau đóng vai trò Query nhằm củng cố quan hệ tương hỗ.

Về mặt toán học, cơ chế cross-attention giữa đặc trưng văn bản F_q và đặc trưng ảnh F_v được mô tả bằng công thức:

$$\text{CrossAttn}(F_q, F_v) = \text{softmax}\left(\frac{F_q F_v^T}{\sqrt{d}}\right) F_v$$

trong đó d là kích thước của không gian đặc trưng. Toán tử softmax đóng vai trò chuẩn hóa trọng số tương tác giữa từng token câu hỏi và từng vùng ảnh, từ đó tạo ra bản đồ attention thể hiện mức độ liên quan giữa từng cặp phần tử.

Cơ chế cross-attention không chỉ cải thiện khả năng căn chỉnh giữa các modality mà còn giúp mô hình học được những mối quan hệ sâu, có tính ngữ nghĩa cao, chẳng hạn như quan hệ thuộc tính (attribute), quan hệ không gian (spatial relation) hoặc quan hệ hành động (action). Nhiều mô hình VQA hiện đại như ViLBERT, LXMERT hay BLIP đều xây dựng module fusion dựa trên cross-attention đa tầng hoặc multi-head cross-attention nhằm khai thác tối đa tính tương tác này.

1.2.6. Unified Vision-Language Models

Các mô hình thị giác–ngôn ngữ (Vision–Language Models, VLMs) hiện đại được phát triển theo hai hướng chính: kiến trúc Modular và kiến trúc Unified. Hướng thứ nhất tổ chức hệ thống thành nhiều thành phần độc lập như bộ mã hóa thị giác, bộ trộn đặc trưng và mô hình sinh ngôn ngữ. Ngược lại, hướng Unified phát triển một kiến trúc hợp nhất, trong đó thông tin hình ảnh và văn bản được xử lý trong cùng một không gian biểu diễn bởi một mô hình duy nhất. Việc so sánh hai hướng tiếp cận này là cần thiết để làm rõ giới hạn tài nguyên, ưu nhược điểm cũng như lựa chọn phù hợp cho phạm vi đề tài.

1.2.6.1. Kiến trúc chung của Unified Vision–Language Models

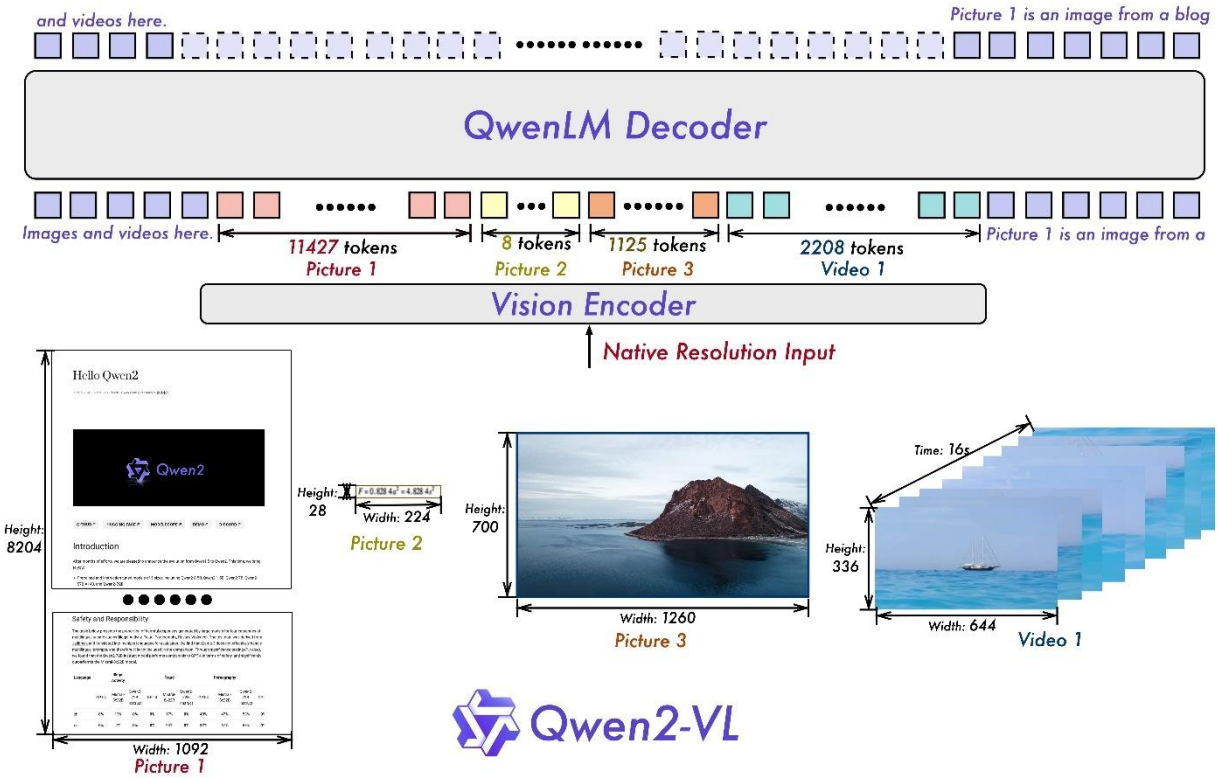
Các mô hình Unified VLMs được thiết kế theo hướng tích hợp sâu đa phương thức, trong đó các đặc trưng hình ảnh được chuyển đổi thành chuỗi "visual tokens" và ghép trực tiếp vào chuỗi văn bản trước khi đưa vào mô hình ngôn ngữ lớn (LLM). Kiến trúc tổng quát thường bao gồm ba thành phần:

- **Vision Encoder:** Thường sử dụng các mô hình như CLIP, SigLIP hoặc các biến thể của Vision Transformer (ViT) để trích xuất đặc trưng hình ảnh dưới dạng embeddings có độ phân giải linh hoạt.
- **Projection Layer:** Lớp chiếu tuyến tính hoặc module cross-attention dùng để chuyển không gian thị giác sang không gian từ vựng của LLM. Phần lớn các VLM hiện đại áp dụng kiến trúc tương tự Q-Former (BLIP-2), trong đó một tập token truy vấn học được giúp kết nối hai miền dữ liệu.
- **Large Language Model (LLM):** Đóng vai trò mô-đun giải mã, xử lý đồng thời visual tokens và text tokens, cho phép mô hình sinh câu trả lời, mô tả cảnh hoặc thực hiện suy luận đa bước.

Cách tổ chức hợp nhất này cho phép thông tin đa phương thức được truyền trực tiếp vào LLM dưới dạng chuỗi token duy nhất, giúp mô hình học được các quan hệ phức tạp giữa ảnh và văn bản thông qua cơ chế attention tự nhiên của transformer.

1.2.6.2. Qwen2-VL – Mô hình Unified VLM hiện đại

Một trong những mô hình đại diện cho hướng Unified là Qwen2-VL [16], được phát triển dựa trên họ mô hình Qwen2. Mô hình này tích hợp thị giác và ngôn ngữ theo thiết kế end-to-end với nhiều cải tiến đáng kể:



Hình 6: Kiến trúc mô hình Qwen2-vl

(Nguồn: Wang et al., 2024 [16])

(1) Kiến trúc thị giác

Qwen2-VL sử dụng một biến thể của Vision Transformer được tối ưu hóa cho hình ảnh có độ phân giải động (dynamic resolution). Điều này cho phép mô hình quan sát chi tiết hình ảnh ở nhiều thang độ mà không làm tăng kích thước token quá mức.

(2) Module chiếu đa phương thức

Dựa trên ý tưởng của Q-Former, Qwen2-VL xây dựng projector dạng cross-attention nhằm chuyển đổi các đặc trưng thị giác thành visual tokens tương thích với không gian embedding của mô hình ngôn ngữ. Bộ projector này cho phép giảm độ mất mát thông tin khi chuyển giữa hai miền, đồng thời duy trì sự ổn định trong quá trình huấn luyện.

(3) Mô hình ngôn ngữ nền

Qwen2-VL sử dụng LLM Qwen2 với hai quy mô chính: 7B và 72B tham số. Mô hình nền này được huấn luyện trên dữ liệu đa ngôn ngữ và đa miền, tạo năng lực sinh ngôn ngữ mạnh và khả năng suy luận hợp lý khi kết hợp cùng tín hiệu thị giác.

(4) Quy trình huấn luyện

Mô hình được huấn luyện theo hai giai đoạn:

- **Giai đoạn 1 – Vision–Language Alignment:** Tối ưu projector và LLM để hiểu được visual tokens.
- **Giai đoạn 2 – Instruction Tuning:** Huấn luyện mô hình với các tập dữ liệu đa phương thức theo dạng hội thoại, giúp nó tuân thủ chỉ dẫn và trả lời tự nhiên hơn.

(5) Cơ chế tích hợp đa phương thức

Thông tin hình ảnh được mã hóa thành visual tokens và được trộn trực tiếp với text tokens trước khi đưa vào LLM. Mô hình sử dụng causal attention mask, đảm bảo tính thống nhất trong quá trình sinh chuỗi, đồng thời cho phép LLM tham chiếu visual tokens tại mọi bước suy luận.

1.2.6.3. So sánh giữa Modular VLMs và Unified VLMs

Để cung cấp cái nhìn tổng quan về hai hướng tiếp cận phổ biến trong các hệ thống Vision–Language Models (VLMs), bảng dưới đây trình bày sự khác biệt cơ bản giữa kiến trúc Modular và Unified (*Bảng 1*).

Bảng 1: So sánh đặc trưng kiến trúc giữa mô hình VLM dạng Modular và Unified

Khía cạnh	**Modular**	**Unified VLM**
Kiến trúc	Tách thành 3 mô-đun độc lập	Hợp nhất end-to-end
Khả năng giải thích	Cao	Thấp
Hỗ trợ tiếng Việt	Tốt	Giới hạn
Dữ liệu huấn luyện	Quy mô trung bình	Tập dữ liệu cực lớn đa ngữ

Số tham số	~500M	7B – 72B
Chi phí fine-tuning	Thấp (LoRA áp dụng lên ít tham số)	Cao (dù dùng LoRA vẫn nặng)
Tốc độ suy luận	Nhanh hơn	Chậm hơn
Phân tích lỗi	Dễ phân tích theo từng mô-đun	Khó phân tích

Nhìn chung, Unified VLMs thường đạt chất lượng cao hơn trong các thiết lập đa phương thức do khai thác hiệu quả cơ chế attention của LLM, nhưng chi phí tính toán và yêu cầu tài nguyên là rất lớn. Ngược lại, Modular VLMs mang lại lợi thế về tốc độ suy luận, khả năng giải thích và sự linh hoạt khi tinh chỉnh.

1.2.7. Parameter-Efficient Fine-Tuning với LoRA

Trong bối cảnh các mô hình ngôn ngữ và mô hình đa phương thức ngày càng có quy mô lớn, việc tinh chỉnh toàn phần (Full Fine-Tuning) trở nên tốn kém cả về chi phí tính toán lẫn bộ nhớ. Phương pháp này yêu cầu cập nhật toàn bộ trọng số của mô hình, đồng nghĩa với việc phải lưu trữ gradients, optimizer states và một bản sao hoàn chỉnh của mô hình cho mỗi tác vụ. Điều này dẫn đến mức tiêu thụ bộ nhớ rất lớn và gây khó khăn trong bối cảnh triển khai thực tế. Đặc biệt, với các mô hình có từ khoảng 100 – 500 triệu tham số, việc nhân bản toàn bộ trọng số cho nhiều tác vụ huấn luyện dễ dẫn đến hàng trăm megabytes dữ liệu cần lưu trữ. Ngoài ra, tinh chỉnh toàn phần còn có nguy cơ gây catastrophic forgetting, khiến mô hình đánh mất tri thức thu được trong giai đoạn tiền huấn luyện.

Để giải quyết những hạn chế này, Low-Rank Adaptation (LoRA [17] được đề xuất như một phương pháp tinh chỉnh hiệu quả về tham số (Parameter-Efficient Fine-Tuning – PEFT). LoRA dựa trên quan sát rằng các cập nhật trọng số trong quá trình tinh chỉnh thực tế có cấu trúc hạng thấp (intrinsic low-rank structure). Thay vì cập nhật trực tiếp ma trận trọng số gốc, LoRA giả định phần thay đổi của trọng số có thể được xấp xỉ bằng tích của hai ma trận hạng thấp. Cụ thể, với ma trận trọng số ban đầu $W \in R^{d \times k}$, LoRA mô hình hóa trọng số sau tinh chỉnh thành:

$$W' = W + \Delta W = W + BA$$

trong đó $B \in R^{d \times r}$ và $A \in R^{r \times k}$ là hai ma trận nhỏ được huấn luyện, với $r \ll min(d, k)$ đóng vai trò như một rank bottleneck. Các trọng số gốc W được giữ cố định trong toàn bộ quá trình huấn luyện, nhờ đó giảm đáng kể chi phí tính toán và hạn chế

hiện tượng quên thảm họa (catastrophic forgetting). Trong quá trình lan truyền xuôi, biểu thức được thực hiện dưới dạng:

$$h = Wx + BAx$$

từ đó cho phép mô hình học được các biểu diễn mới thông qua các ma trận hạng thấp mà không cần thay đổi cấu trúc gốc.

LoRA thường được chèn vào các module tuyến tính trong kiến trúc Transformer, đặc biệt là các phép chiếu Query, Key và Value trong cơ chế Self-Attention [17]. Việc lựa chọn các module này giúp đảm bảo khả năng biểu đạt của mô hình mà không làm tăng đáng kể chi phí suy luận. Bộ siêu tham số của LoRA bao gồm: rank r, quyết định mức độ biểu diễn của cập nhật trọng số; và scaling factor α, đóng vai trò điều chỉnh độ lớn của tập cập nhật thông qua tỉ lệ α/r. Trong nhiều thực nghiệm, các giá trị nhỏ như r = 8-16 cho các mô hình encoder và r = 16-32 cho các mô hình encoder-decoder hoặc vision-language models thường mang lại hiệu năng tốt với chi phí tối thiểu [17].

Một ưu điểm nổi bật của LoRA là khả năng giảm số lượng tham số cần huấn luyện xuống chỉ còn khoảng 0.1–0.3% tổng số tham số gốc của mô hình, giúp tiết kiệm đáng kể bộ nhớ và dung lượng lưu trữ [17]. Kết quả thực nghiệm trong các nghiên cứu cho thấy LoRA thường đạt từ 95–99% hiệu năng của phương pháp tinh chỉnh toàn phần, trong khi chi phí huấn luyện giảm mạnh và giúp mô hình giữ lại tri thức ngôn ngữ tích lũy từ giai đoạn tiền huấn luyện [17].

Đối với các hệ thống VQA tiếng Việt, LoRA mang lại nhiều lợi ích quan trọng. Thứ nhất, việc giữ nguyên trọng số nền của mô hình ngôn ngữ giúp bảo toàn tri thức ngôn ngữ tiếng Việt vốn được học từ lượng dữ liệu lớn, đồng thời chỉ cập nhật những thành phần cần thiết để thích ứng với nhiệm vụ liên quan đến hình ảnh. Thứ hai, LoRA cho phép xây dựng mô hình đa nhiệm (multi-task) bằng cách chia sẻ backbone cố định và chỉ thay đổi các adapter LoRA cho từng tác vụ, phù hợp với môi trường nghiên cứu và triển khai giới hạn tài nguyên. Cuối cùng, chi phí huấn luyện thấp giúp các mô hình VQA có thể được triển khai hiệu quả trong thực tế, đặc biệt ở các hệ thống yêu cầu cập nhật thường xuyên mà không muốn huấn luyện lại toàn bộ mô hình.

1.2.8. Các chỉ số đánh giá (Evaluation Metrics)

Trong các hệ thống Visual Question Answering (VQA), việc đánh giá chất lượng câu trả lời sinh tự động là một thách thức đáng kể do tính chất mở (open-ended) và đa dạng của ngôn ngữ tự nhiên. Một câu trả lời có thể đúng về mặt ngữ nghĩa mặc dù hoàn toàn không trùng khớp về mặt từ vựng, hoặc ngược lại. Do vậy, các chỉ số đánh giá cần xem xét đồng thời hai khía cạnh: sự trùng khớp bề mặt (lexical similarity) và tương đồng ngữ nghĩa (semantic similarity). Trong phạm vi luận văn, năm chỉ số phổ biến BLEU, ROUGE, CIDEr, BERTScore và Exact Match (EM) được sử dụng để phản ánh những góc nhìn khác nhau về chất lượng mô hình.

1.2.8.1. BLEU

BLEU (Bilingual Evaluation Understudy) [18] là chỉ số đánh giá kinh điển trong dịch máy, hiện được sử dụng rộng rãi trong các bài toán sinh ngôn ngữ. BLEU đo mức độ tương đồng giữa câu trả lời dự đoán c và tập câu tham chiếu S thông qua precision của n-gram, kết hợp với cơ chế phạt câu ngắn (brevity penalty – BP).

Công thức tổng quát của BLEU được định nghĩa như sau:

$$BLUE = BP \times \exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

Trong đó:

- p_n là precision của n-gram bậc nnn:

$$p_n = \frac{\sum_{ngram \in c} min(Count_c(ngram), max_{s \in S} Count_n(ngram))}{\sum_{ngram \in c} Count_c(ngram)}$$

- w_n là trọng số của n-gram (thường $w_n = 1/N$).
- BP là hệ số phạt độ dài:

$$BP = \begin{cases} 1, & nếu\ c > r \\ e^{\left(1-\frac{r}{c}\right)}, & nếu\ c \leq r \end{cases}$$

với c là độ dài câu dự đoán, r là độ dài câu tham chiếu tốt nhất (closest reference length).

BLEU được thiết kế để ưu tiên sự chính xác trong việc tái tạo cấu trúc n-gram, đồng thời hạn chế việc mô hình sinh câu quá ngắn nhằm "ăn gian" precision.

1.2.8.2. ROUGE

ROUGE [19] (Recall-Oriented Understudy for Gisting Evaluation) là nhóm chỉ số thường được sử dụng trong tóm tắt văn bản, nhấn mạnh yếu tố recall. Trong VQA, ROUGE giúp đánh giá mức độ bao phủ nội dung trong câu trả lời dự đoán.

(1) ROUGE-N

ROUGE-N đo mức độ trùng khớp của n-gram giữa câu tham chiếu và câu dự đoán. Công thức:

$$ROUGE_N = \frac{\sum_{gram_N \in reference} Count_{match}(gram_N)}{\sum_{gram\ N \in reference} Count(gram_N)}$$

Trong đó:

- $Count_{match}$: số lần n-gram xuất hiện trong cả câu tham chiếu và câu dự đoán.
- Chỉ số phổ biến nhất là ROUGE-1 (unigram) và ROUGE-2 (bigram).

(2) ROUGE-L

ROUGE-L sử dụng Longest Common Subsequence (LCS) để đánh giá độ tương đồng theo trình tự từ – cho phép bắt được sự tương đồng ngữ nghĩa ngay cả khi các từ không liên tiếp.

Giả sử LCS giữa câu tham chiếu S và câu dự đoán c có độ dài $LCS(S, c)$.Ta có:

$$R_{LCS} = \frac{LCS(S, c)}{|S|}$$

$$P_{LCS} = \frac{LCS(S, c)}{|c|}$$

Chỉ số ROUGE-L được biểu diễn dưới dạng F-score:

$$ROUGE_L = \frac{(1 + \beta^2) \times R_{LCS} \times P_{LCS}}{\beta^2 P_{LCS} + R_{LCS}}$$

Với $\beta = 1$, ROUGE-L trở thành F1-score của LCS.

1.2.8.3. CIDEr

CIDEr (Consensus-based Image Description Evaluation) [20] được thiết kế dành riêng cho đánh giá mô tả ảnh và ngôn ngữ đa phương thức. Chỉ số này nhấn mạnh các từ mang tính nội dung, sử dụng vector đặc trưng TF–IDF và đo độ tương đồng bằng cosine similarity.

Giả sử với mỗi câu dự đoán c và câu tham chiếu s_i, ta biểu diễn TF–IDF của n-gram bậc n dưới dạng vector $g_n(c)$ và $g_n(s_i)$. CIDEr-n được tính như sau:

$$CIDEr_n = \frac{1}{M} \sum_{i=1}^{M} \frac{g_n(c) \times g_n(s_i)}{\|g_n(c)\| \|g_n(s_i)\|}$$

Bộ chỉ số tổng hợp:

$$CIDEr = \sum_{n=1}^{4} w_n \times CIDEr_n$$

Trong đó thường chọn $w_n = 1/4$ cho 4-gram.

CIDEr có ưu điểm vượt trội so với BLEU và ROUGE nhờ cân nhắc được tần suất quan trọng của từ trong toàn bộ tập dữ liệu thông qua IDF, giúp đánh giá tốt hơn cho các tác vụ mô tả ảnh và trả lời câu hỏi dựa trên ảnh.

1.2.8.4. BERTScore

BERTScore [21] được xây dựng nhằm khắc phục hạn chế của BLEU/ROUGE trong việc đánh giá sự tương đồng ngữ nghĩa. Thay vì so khớp n-gram, BERTScore sử dụng embedding ngữ cảnh từ mô hình Transformer (như BERT, RoBERTa) và tính toán cosine similarity giữa từng từ trong câu dự đoán và câu tham chiếu.

Cho câu dự đoán $c = \{c_1, \ldots, c_m\}$ và câu tham chiếu $s = \{s_1, \ldots, s_n\}$ với embedding tương ứng v_{c_i} và v_{s_j}, BERTScore được tính như sau:

(1) Precision:

$$\text{Precision} = \frac{1}{\text{m}} \sum_{\text{i}=1}^{\text{m}} \max_{\text{j}} \cos\left(\text{v}_{\text{c}_\text{i}}, \text{v}_{\text{s}_\text{j}}\right)$$

(2) Recall:

$$Recall = \frac{1}{n} \sum_{j=1}^{n} \max_{i} cos\left(v_{s_j}, v_{c_i}\right)$$

(3) F1-score:

$$BERTScore = F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Chỉ số này phản ánh mức độ tương đồng ngữ nghĩa sâu, không phụ thuộc vào hình thức n-gram, đặc biệt phù hợp với các mô hình ngôn ngữ hiện đại.

1.2.8.5. Exact Match (EM)

Exact Match (EM) [22] là thước đo đơn giản nhưng quan trọng trong nhiều bài toán hỏi đáp. EM kiểm tra liệu câu trả lời dự đoán có khớp hoàn toàn với câu tham chiếu hay không (thường sau khi chuẩn hóa loại bỏ dấu câu, chữ hoa/thường).

$$EM = \begin{cases} 1, & nếu\ c = S \\ 0, & ngược\ lại \end{cases}$$

Chỉ số EM phản ánh khả năng mô hình tạo ra câu trả lời chính xác tuyệt đối, đặc biệt quan trọng trong những tập dữ liệu có cấu trúc đóng (như câu trả lời dạng "yes/no", số, hoặc tên riêng).

1.2.9. Các thư viện và công cụ

Trong quá trình nghiên cứu và triển khai mô hình VQA, hệ thống sử dụng nhiều thư viện và công cụ khác nhau để hỗ trợ xử lý dữ liệu, huấn luyện và đánh giá mô hình. Các thành phần chính gồm:

(1) Các framework học sâu

PyTorch 2.0 đây là nền tảng học sâu chủ đạo được sử dụng trong hệ thống nhờ khả năng hỗ trợ đồ thị tính toán động, hiệu năng cao và cộng đồng phát triển mạnh mẽ. Phiên bản 2.0 tích hợp TorchDynamo và các cải tiến tối ưu hóa giúp tăng tốc quá trình huấn luyện và suy luận.

Transformers (Hugging Face) thư viện cung cấp các mô hình tiền huấn luyện quy mô lớn và các thành phần xử lý ngôn ngữ – thị giác tích hợp sẵn. Việc sử dụng Transformers giúp rút ngắn thời gian xây dựng pipeline, đồng thời đảm bảo tính tương thích với các mô hình Vision-Language hiện đại như BLIP, Qwen-VL hay LLaVA.

(2) Thị giác máy tính

Pillow thư viện xử lý ảnh cơ bản được dùng để tiền xử lý dữ liệu đầu vào, bao gồm chuyển đổi định dạng, thay đổi kích thước, chuẩn hóa và các thao tác thao tác pixel-level cần thiết trước khi đưa ảnh vào mô hình.

qwen-vl-utils là bộ công cụ hỗ trợ riêng cho dòng mô hình Qwen-VL, cung cấp các hàm xử lý ảnh, trích xuất đặc trưng thị giác và chuẩn hóa dữ liệu đa phương thức theo đúng định dạng của mô hình. Công cụ này đảm bảo tính nhất quán giữa quy trình tiền xử lý và kiến trúc Vision-Language mà hệ thống sử dụng.

(3) Xử lý ngôn ngữ tự nhiên (NLP)

PyVi thư viện được sử dụng cho tác vụ tách từ và chuẩn hóa văn bản tiếng Việt. Việc tách từ đặc biệt quan trọng trong quá trình đánh giá tự động, giúp các thước đo dựa trên n-gram (như BLEU hoặc ROUGE) phản ánh chất lượng mô hình một cách chính xác hơn đối với ngôn ngữ tiếng Việt.

(4) Các tiện ích cho huấn luyện

PEFT (Parameter-Efficient Fine-Tuning) là framework triển khai các phương pháp tinh chỉnh hiệu quả tham số, trong đó quan trọng nhất là LoRA. PEFT cho phép huấn luyện các mô hình lớn với chi phí bộ nhớ thấp, đồng thời hạn chế hiện tượng quên thảm họa trong quá trình tinh chỉnh.

Bitsandbytes thư viện cung cấp cơ chế lượng tử hóa 4-bit và tối ưu hóa dựa trên 8-bit, giúp giảm đáng kể dung lượng bộ nhớ và tăng hiệu quả khi huấn luyện các mô hình lớn. Việc kết hợp bitsandbytes với LoRA trong PEFT tạo thành quy trình QLoRA, phù hợp cho môi trường có tài nguyên hạn chế như Google Colab, kaggle hoặc GPU phổ thông.

(5) Các công cụ đánh giá

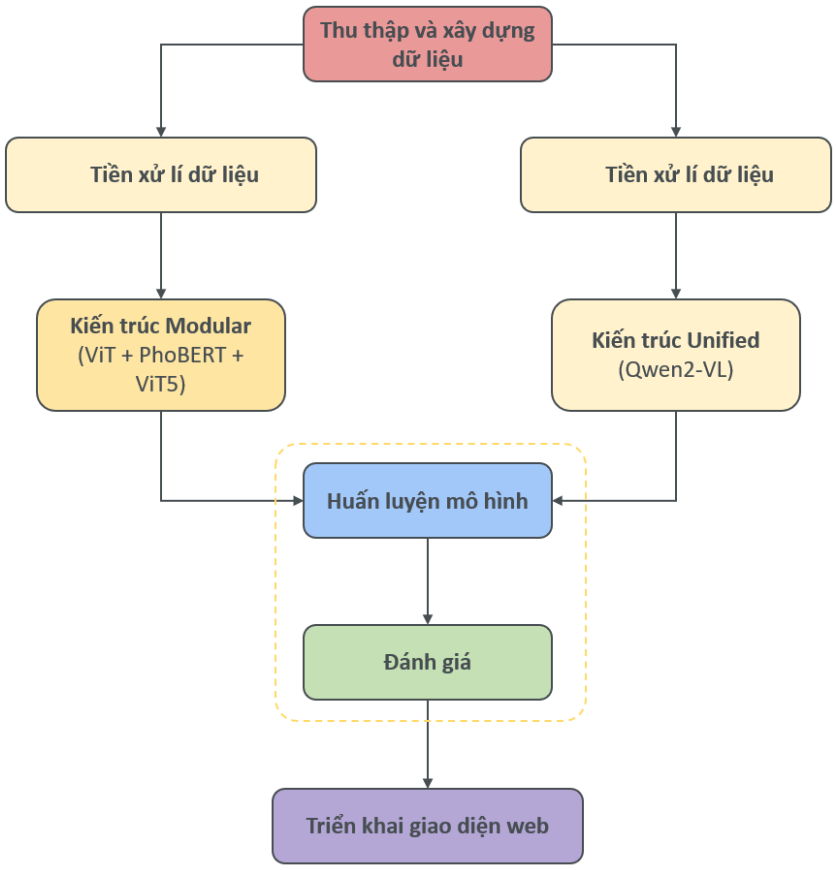
Hệ thống sử dụng nhiều thư viện để tính toán các metrics đánh giá chất lượng sinh ngôn ngữ, bao gồm **pycocoevalcap (CIDEr)**, **bert-score (BERTScore)**, **rouge-score (ROUGE)** và **NLTK (BLEU)**. Việc kết hợp đa dạng các thước đo này giúp mô hình được đánh giá một cách toàn diện hơn, bao phủ cả mức độ trùng khớp bề mặt (lexical similarity), mức độ tương đồng ngữ nghĩa (semantic similarity) và khả năng chính xác tổng thể của câu trả lời.

CHƯƠNG 2. THIẾT KẾ VÀ CÀI ĐẶT GIẢI PHÁP

2.1. Thiết kế hệ thống

2.1.1. Tổng quan kiến trúc hệ thống

Hệ thống Visual Question Answering (VQA) được thiết kế theo kiến trúc module hóa, cho phép vận hành song song hai hướng mô hình: (i) mô hình Modular với các khối xử lý độc lập và (ii) mô hình Unified hoạt động theo cơ chế end-to-end. Mặc dù cấu trúc chi tiết của hai hướng tiếp cận khác nhau, quy trình tổng thể của hệ thống vẫn bao gồm ba thành phần chính: tiền xử lý dữ liệu, pipeline mô hình và giai đoạn huấn luyện – suy luận (*Hình 7*).



Hình 7: Tổng quan kiến trúc hệ thống

(1) Tiền xử lí dữ liệu

Trong giai đoạn tiền xử lý, ảnh và câu hỏi tiếng Việt được xử lý theo hai pipeline tách biệt.

Đối với mô hình Modular, câu hỏi được làm sạch, phân đoạn từ tiếng Việt và chuẩn hóa câu trả lời; ảnh được chuyển đổi và chuẩn hóa theo định dạng yêu cầu của ViT. Đối với mô hình Unified, dữ liệu được chuẩn hóa nhẹ hơn, chủ yếu tập trung vào việc định dạng đầu vào theo cấu trúc đa phương thức của Qwen2-VL (ảnh + câu hỏi dưới dạng prompt).

Các mẫu sau xử lý được gom nhóm theo từng ảnh để bảo đảm tính nhất quán và dễ quản lý trong các bước tiếp theo.

(2) Pipeline mô hình

Pipeline của hai mô hình được tổ chức theo hai cơ chế khác nhau. Trong mô hình Modular, ảnh được mã hóa bằng ViT, câu hỏi được mã hóa bằng PhoBERT, sau đó đặc trưng từ hai miền được kết hợp thông qua mô-đun fusion (cross-attention kết hợp gating) và chiếu sang không gian ẩn của bộ giải mã ViT5. Ngược lại, mô hình Unified sử dụng trực tiếp Qwen2-VL để tiếp nhận đồng thời ảnh và câu hỏi trong một pipeline duy nhất, nơi toàn bộ quá trình mã hóa và hợp nhất được xử lý nội bộ bởi mô hình.

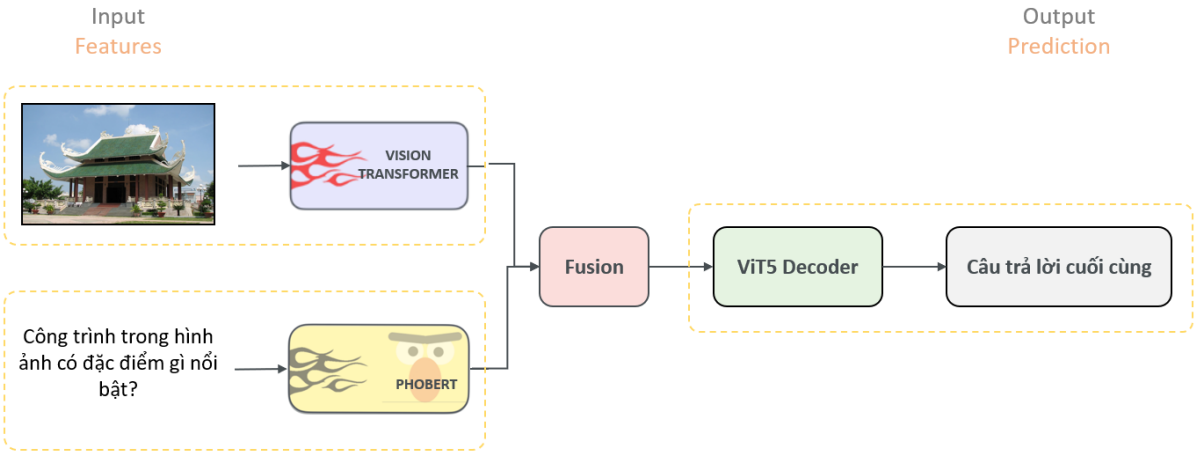
(3) Huấn luyện và suy luận

Trong giai đoạn huấn luyện, mô hình Modular sinh câu trả lời bằng ViT5 theo cơ chế tự hồi quy, với gradient chỉ cập nhật vào các tham số LoRA và các lớp tùy chỉnh như fusion và projection.

Mô hình Unified tối ưu trực tiếp tham số của Qwen2-VL trên định dạng đầu vào đa phương thức.

Cả hai pipeline đều sử dụng hàm mất mát sinh chuỗi và chia sẻ cùng bộ chỉ số đánh giá trên tập kiểm thử nhằm bảo đảm tính nhất quán trong so sánh.

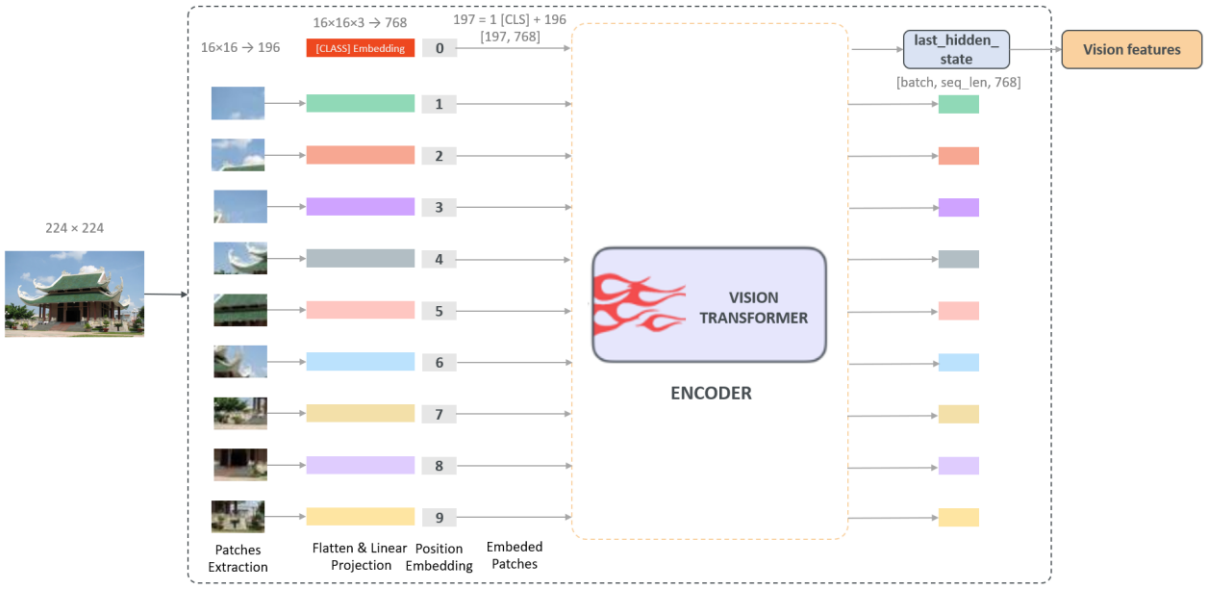
2.1.2. Thiết kế kiến trúc Modular (ViT+PhoBERT+ViT5)



Hình 8: Kiến trúc Modular với ViT, PhoBERT và ViT5

Hình 8 mô tả kiến trúc Modular được xây dựng từ ba mô-đun độc lập, tương ứng với ba loại thông tin: đặc trưng ảnh, đặc trưng câu hỏi và mô-đun sinh câu trả lời. Cách tổ chức theo mô-đun cho phép quá trình huấn luyện linh hoạt hơn, dễ dàng thay thế hoặc tinh chỉnh từng thành phần, đồng thời phù hợp với bối cảnh bài toán VQA tiếng Việt, nơi các mô hình đa phương thức chuyên biệt chưa thực sự phổ biến.

2.1.2.1. Trích xuất đặc trưng ảnh bằng ViT



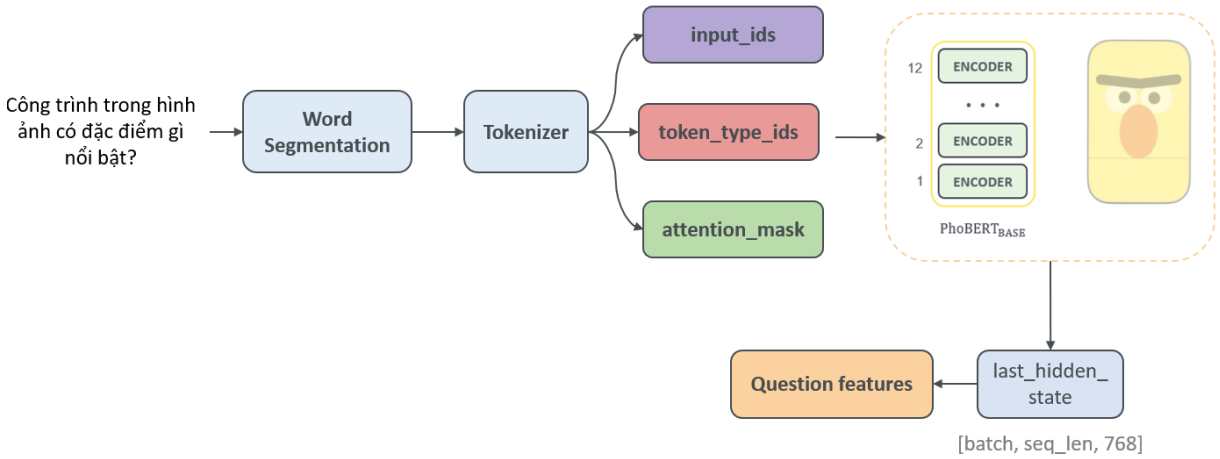
Hình 9: Kiến trúc Vision Transformer (ViT) cho xử lý ảnh đầu vào

Hình 9 minh họa kiến trúc Vision Transformer (ViT) được sử dụng để trích xuất đặc trưng thị giác từ hình ảnh. Hình ảnh gốc có kích thước 224×224 được chia thành 196 patch có kích thước 16×16. Mỗi patch sau đó được làm phẳng và ánh xạ vào không gian ẩn với kích thước 768 thông qua một lớp Linear Projection. Một token đặc biệt [CLS] được thêm vào đầu chuỗi nhằm đại diện cho toàn bộ thông tin hình ảnh. Các embedding này tiếp tục được cộng với positional embedding để bổ sung thông tin vị trí trước khi đưa vào chuỗi Transformer Encoder. Cấu trúc này cho phép mô hình học các mối quan hệ không gian và ngữ nghĩa toàn cục giữa các patch của ảnh.

Trong kiến trúc Modular, toàn bộ trọng số của backbone ViT được giữ cố định; chỉ các tham số LoRA gắn vào các lớp attention được cập nhật trong quá trình huấn luyện. Cách tiếp cận này giúp giảm đáng kể chi phí tính toán và bộ nhớ, đồng thời vẫn duy trì chất lượng biểu diễn thị giác của mô hình gốc.

Đầu ra của ViT là tensor last_hidden_state có kích thước [batch_size, seq_len, 768], trong đó seq_len = 197, bao gồm 1 token [CLS] và 196 patch embedding. Toàn bộ chuỗi embedding này sau đó được sử dụng làm đặc trưng thị giác đầu vào cho mô-đun Fusion, đóng vai trò quan trọng trong việc kết hợp thông tin đa phương thức.

2.1.2.2. Trích xuất đặc trưng câu hỏi bằng PhoBERT



Hình 10: Kiến trúc Question Encoder sử dụng PhoBERT

Hình 10 minh họa mô-đun Question Encoder, trong đó PhoBERT được sử dụng để mã hóa câu hỏi tiếng Việt thành các đặc trưng ngôn ngữ. Sau khi thực hiện các bước làm sạch và phân đoạn từ, câu hỏi được đưa vào bộ tokenizer của PhoBERT để sinh ra ba thành phần đầu vào:

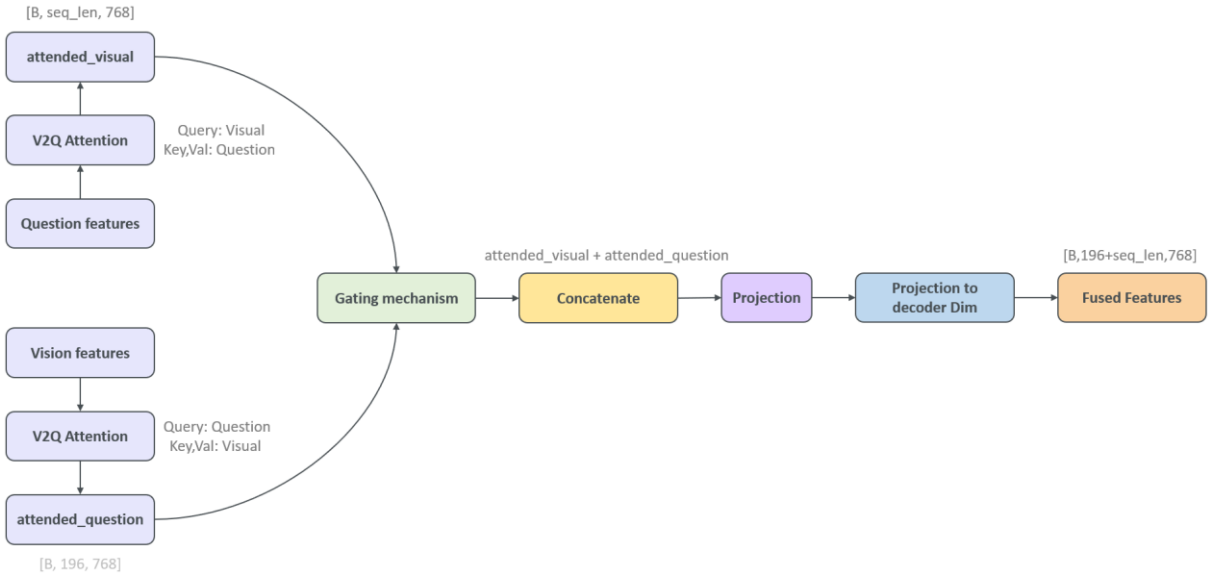
- **input_ids:** chuỗi token đã được mã hóa theo chuẩn BPE;
- **token_type_ids:** mã phân tách đoạn câu (đối với câu hỏi đơn, giá trị này thường đồng nhất);
- **attention_mask:** mặt nạ xác định các vị trí cần được mô hình chú ý.

Ba thành phần này sau đó được đưa qua 12 lớp Transformer Encoder của PhoBERT. Thông qua cơ chế self-attention, mô hình học được các mối quan hệ cú pháp và ngữ nghĩa giữa các token trong câu hỏi.

Đầu ra của PhoBERT là last_hidden_state với kích thước [batch_size, seq_len, 768]. Đây chính là Question Features, bao gồm embedding của từng token sau khi đi qua toàn bộ chuỗi encoder. Các đặc trưng này đóng vai trò quan trọng trong việc kết hợp với đặc trưng thị giác trong mô-đun Fusion.

Tương tự như trong mô-đun ViT, trọng số backbone của PhoBERT được giữ cố định; chỉ các tham số LoRA gắn vào các lớp attention được cập nhật trong quá trình huấn luyện. Thiết kế này giúp giảm đáng kể chi phí tính toán và bộ nhớ, đồng thời vẫn duy trì chất lượng biểu diễn ngôn ngữ phù hợp cho bài toán VQA tiếng Việt.

2.1.2.3. Fusion Module



Hình 11: Module Fusion: Kết hợp Vision và Question Features

Hình 11 minh họa mô-đun Fusion, thành phần trung tâm trong kiến trúc Modular. Tại mô-đun này, embedding hình ảnh từ ViT và embedding câu hỏi từ PhoBERT được kết hợp thông qua một khối cross-attention hai chiều. Cơ chế này cho phép:

- Đặc trưng hình ảnh truy vấn các đặc trưng ngôn ngữ, nhằm xác định thông tin liên quan trong câu hỏi.
- Đặc trưng câu hỏi truy vấn các đặc trưng hình ảnh, tập trung vào những vùng trực quan phù hợp với ngữ cảnh câu hỏi.

Sự tương tác hai chiều này tăng mức độ hòa trộn giữa hai modality, giúp mô hình học được các quan hệ phụ thuộc sâu và giàu ngữ nghĩa.

Sau khối cross-attention, một gating layer được áp dụng để điều chỉnh mức độ đóng góp tương đối giữa hai nguồn thông tin, qua đó nâng cao tính ổn định của biểu diễn hợp nhất. Hai embedding đã được điều biến bởi gating layer được ghép nối (concatenate) thành một vector đa phương thức. Vector này tiếp tục được đưa qua một lớp projection, nhằm chuẩn hóa và biến đổi về không gian biểu diễn thống nhất. Một bước chiếu bổ sung được thực hiện để ánh xạ biểu diễn hợp nhất về đúng kích thước ẩn (decoder hidden size) của ViT5, đảm bảo khả năng tương thích với bộ giải mã.

Đầu ra cuối cùng của mô-đun Fusion là embedding đa phương thức đã được chuẩn hóa và điều chỉnh kích thước, sẵn sàng làm đầu vào cho ViT5 Decoder trong nhiệm vụ sinh câu trả lời.

2.1.2.4. Mô-đun sinh câu trả lời bằng ViT5

Mô-đun Answer Generation sử dụng ViT5 đóng vai trò là bộ giải mã ngôn ngữ. Biểu diễn đa phương thức từ mô-đun Fusion, đã được chuẩn hóa và điều chỉnh về kích

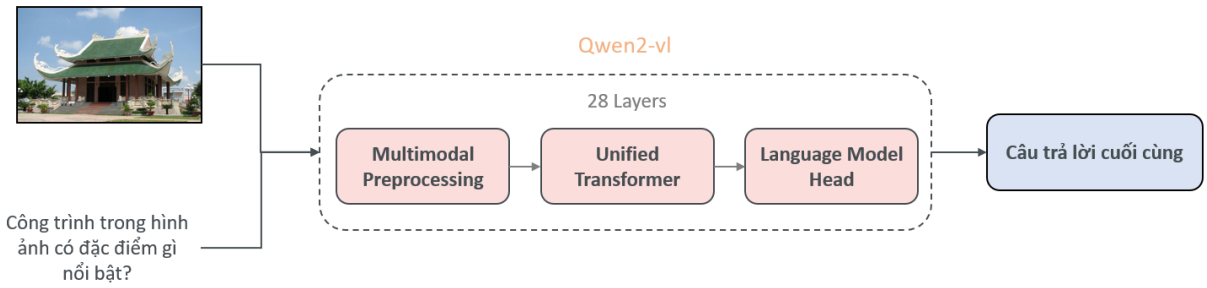
thước ẩn, được sử dụng làm bộ nhớ để decoder truy vấn thông qua các lớp cross-attention.

ViT5 sinh câu trả lời theo cơ chế tự hồi quy: tại mỗi bước, mô hình dự đoán token tiếp theo dựa trên các token đã sinh trước đó kết hợp với thông tin đa phương thức từ Fusion. Các lớp attention trong decoder đảm bảo rằng quá trình truy vấn này diễn ra nhất quán với đặc trưng hình ảnh và câu hỏi.

Trong quá trình huấn luyện, backbone của ViT5 được giữ cố định; chỉ các tham số LoRA gắn vào các lớp attention được cập nhật. Thiết kế này giúp giảm chi phí huấn luyện đồng thời vẫn duy trì khả năng thích ứng với nhiệm vụ VQA tiếng Việt.

Đầu ra cuối cùng là chuỗi token tiếng Việt đã được giải mã, biểu diễn câu trả lời phù hợp với thông tin chứa trong ảnh và câu hỏi.

2.1.3. Thiết kế kiến trúc Unified (Qwen2-VL)



Hình 12: Kiến trúc Unified với Qwen2-VL

Qwen2-VL áp dụng kiến trúc Unified, trong đó thông tin hình ảnh và văn bản được xử lý trong cùng một không gian biểu diễn, thay vì tách thành các encoder độc lập như các mô hình truyền thống. Cách tiếp cận này cho phép mô hình học trực tiếp các quan hệ tương tác giữa hai modality, đồng thời tối ưu hóa hiệu quả tính toán.

Kiến trúc của Qwen2-VL gồm ba thành phần chính (*Hình 12*). Multimodal Preprocessing chịu trách nhiệm chuẩn hóa và mã hóa dữ liệu đầu vào. Hình ảnh được resize, chuyển sang RGB và đưa vào Vision Encoder dựa trên Vision Transformer. Tại đây, ảnh được phân tách thành các visual tokens với số lượng linh hoạt tùy theo độ phân giải (thông thường khoảng 200–300 tokens với kích thước 336×336). Câu hỏi được token hóa và ánh xạ thành các vector embedding, đồng thời được gắn vào một template hội thoại chuẩn hóa, sử dụng các special tokens để phân biệt vai trò (người dùng – trợ lý) và loại nội dung (ảnh – văn bản), giúp mô hình duy trì cấu trúc tương tác ngôn ngữ thống nhất.

Unified Transformer (28 layers) tiếp nhận toàn bộ embedding từ ảnh và văn bản thông qua việc nối chuỗi visual tokens và text tokens thành một chuỗi hợp nhất. Chuỗi này được xử lý xuyên suốt qua tất cả các lớp Transformer, trong đó cơ chế self-attention cho phép các token tương tác trực tiếp, hình thành biểu diễn cross-modal. Đồng thời,

positional encoding được sử dụng để bảo toàn trật tự và thông tin vị trí của các token trong chuỗi đa phương thức.

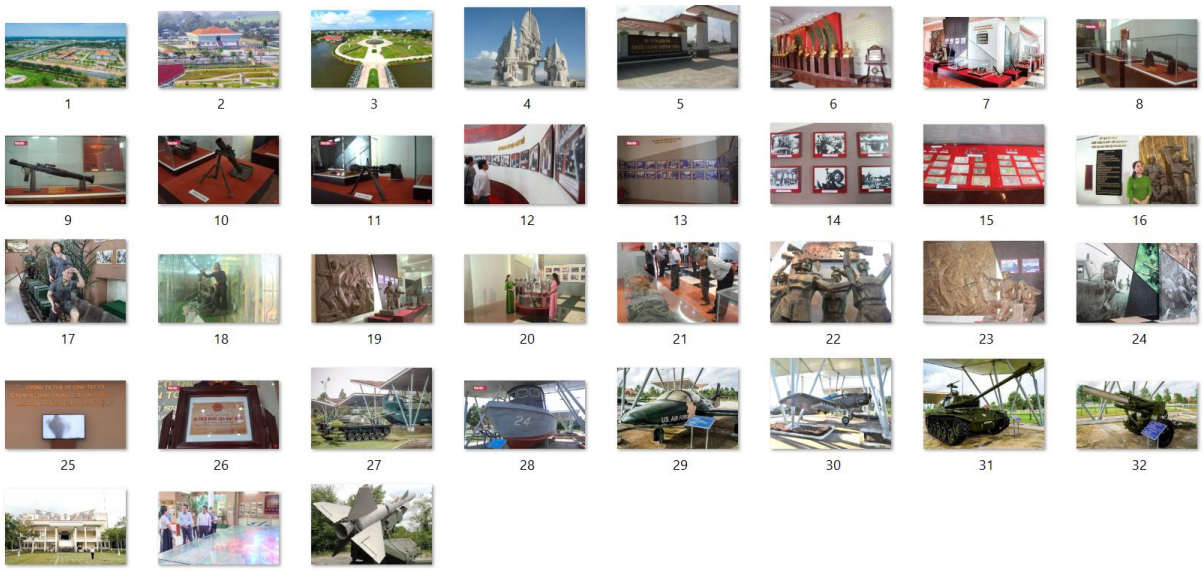
Cuối cùng, Language Model Head sinh câu trả lời theo cơ chế autoregressive, dựa trên biểu diễn ẩn sau Unified Transformer. Các hidden states được chiếu lên không gian từ vựng để dự đoán token kế tiếp. Kiến trúc Unified không chỉ mang lại khả năng tích hợp sâu giữa hai modality, mà còn giảm chi phí tính toán nhờ sử dụng duy nhất một backbone Transformer, đồng thời tạo điều kiện thuận lợi cho việc mở rộng sang các modality khác khi cần thiết.

2.2. Cài đặt giải pháp

2.2.1. Thu thập và xây dựng dữ liệu

2.2.1.1. Thu thập dữ liệu

Quá trình thu thập dữ liệu nhằm xây dựng nguồn tư liệu hình ảnh và thông tin phục vụ cho hệ thống Visual Question Answering (VQA) về di sản văn hóa cách mạng vùng Tây Nam Bộ. Nguồn dữ liệu hình ảnh được tổng hợp từ nhiều kênh khác nhau nhằm đảm bảo tính đa dạng, đại diện và phản ánh chính xác hiện trạng các di tích.



Hình 13: Minh họa dữ liệu hình ảnh

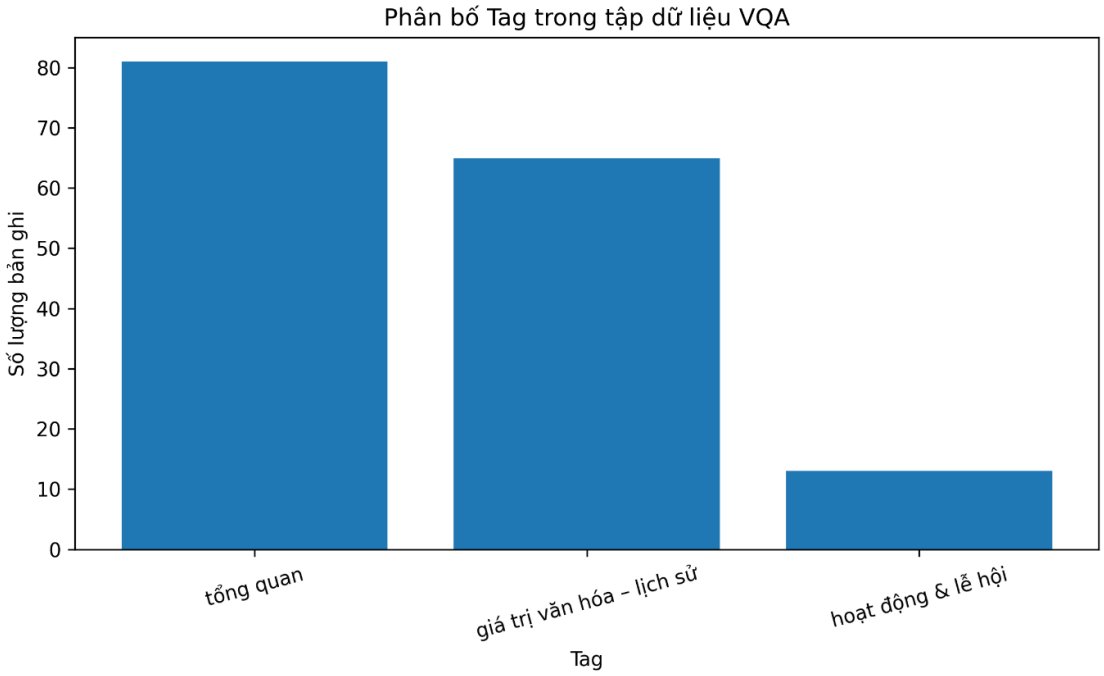
Cụ thể, hình ảnh được thu thập từ các nguồn sau:

- Các cổng thông tin điện tử của tỉnh, sở văn hóa và trang chính thức giới thiệu di tích.
- Bài viết báo chí và tài liệu truyền thông địa phương.
- Video tư liệu, từ đó trích xuất thủ công các khung hình liên quan đến di tích.
- Các trang web mở, công khai với chất lượng hình ảnh phù hợp cho bài toán thị giác máy tính.

Trong quá trình thu thập, dữ liệu được kiểm tra kỹ lưỡng để loại bỏ các hình ảnh trùng lặp, mờ, thiếu thông tin hoặc không phản ánh đúng nội dung di tích. Tên gọi và vị trí hành chính của mỗi di tích được thống nhất trước và sau khi thay đổi, nhằm đảm bảo tính nhất quán cho tập dữ liệu. Tổng số di tích được thu thập là 81, với 1.226 hình ảnh.

2.2.1.2. Xây dựng tập dữ liệu

Sau khi thu thập, dữ liệu được tổ chức thành tập dữ liệu VQA. Mỗi di tích được phân loại thành nhiều bản ghi theo các tag khác nhau, nhằm phản ánh các khía cạnh: tổng quan, giá trị văn hóa – lịch sử, và hoạt động – lễ hội.



Hình 14: Phân bố Tag trong tập dữ liệu VQA

Quy trình tạo cặp Question–Answer (Q-A) được chuẩn hóa như sau:

- Mã hóa hình ảnh sang định dạng Base64;
- Gửi hình ảnh cùng mô tả ngắn và tag tới mô hình ngôn ngữ (GPT-4o-mini) để sinh câu hỏi và câu trả lời phù hợp;
- Mỗi cặp Q-A dựa duy nhất trên quan sát trực tiếp của hình ảnh, không suy diễn hay bổ sung thông tin bên ngoài;
- Kết quả được lưu dưới dạng JSON và tích hợp trực tiếp vào tập dữ liệu, đảm bảo đồng nhất cấu trúc và sẵn sàng cho huấn luyện.

Tổng số cặp Q-A trong tập dữ liệu là 5.631, trung bình mỗi hình ảnh có từ 4–5 cặp Q-A, tương ứng với các khía cạnh khác nhau của di tích. Hình 15 minh họa cấu trúc bản ghi ví dụ của di tích Địa điểm Chiến thắng Chi khu Ngã Năm, thuộc tag “Tổng quan”, bao gồm mô tả di tích, hình ảnh và bộ câu hỏi – trả lời liên quan.

```
{
    "name": "Di tích lịch sử Địa điểm Chiến thắng Chi khu Ngã Năm",
    "location_before": "phường 1, thị xã Ngã Năm, tỉnh Sóc Trăng",
    "location_after": "phường Ngã Năm, thành phố Cần Thơ",
    "tag": "tổng quan",
    "description": "Hình ảnh di tích Địa điểm Chiến thắng Chi khu Ngã Năm một công trình tưởng niệm [...] mang đậm giá trị lịch sử và văn hóa.",
    "images": [
        {
            "image": "images/soc_trang/dia_diem_chien_thang_chi_khu_nga_nam/1.jpg",
            "qas": [
                {
                    "question": "Bức ảnh thể hiện hình ảnh của di tích nào?",
                    "answer": "Bức ảnh thể hiện di tích lịch sử Địa điểm Chiến thắng Chi khu Ngã Năm."
                },
                {
                    "question": "Các yếu tố kiến trúc nào nổi bật trong bức ảnh?",
                    "answer": "Các yếu tố kiến trúc nổi bật bao gồm tượng đài lớn, bậc thang và các chi tiết trang trí ở xung quanh."
                },
                {
                    "question": "Môi trường xung quanh di tích được mô tả như thế nào?",
                    "answer": "Môi trường xung quanh di tích có cây xanh, sân vườn và các công trình hỗ trợ như đèn chiếu sáng."
                },
                {
                    "question": "Góc nhìn của bức ảnh là từ đâu?",
                    "answer": "Bức ảnh được chụp từ phía trước của công trình, cho thấy toàn cảnh."
                }
            ]
        }
    ]
}
```


Hình 15: Cấu trúc bản ghi di tích trong tập dữ liệu

Nhờ cách tổ chức này, tập dữ liệu cuối cùng đảm bảo cấu trúc đồng nhất, bao phủ đa dạng các di tích và khía cạnh quan sát, đồng thời sẵn sàng phục vụ cho quá trình huấn luyện mô hình VQA.

2.2.2. Chuẩn bị dữ liệu

Sau khi xây dựng tập dữ liệu VQA, bước tiếp theo là chuẩn bị dữ liệu cho quá trình huấn luyện và đánh giá mô hình. Tập dữ liệu được tổ chức theo hình ảnh, nghĩa là tất cả các cặp Q-A liên quan đến cùng một hình ảnh được gộp chung để tránh việc rò rỉ thông tin giữa các tập. Việc này giúp đảm bảo rằng mô hình không "nhìn thấy" trước các hình ảnh đã xuất hiện trong tập huấn luyện khi đánh giá trên tập kiểm thử, qua đó giảm nguy cơ data leakage.

Tập dữ liệu được chia theo tỉ lệ 80% – 10% – 10% cho các tập huấn luyện – kiểm định – kiểm thử. Cụ thể:

- **Tập huấn luyện (train) (80%):** Sử dụng để huấn luyện mô hình và điều chỉnh các tham số học;
- **Tập kiểm định (validation) (10%):** Dùng để đánh giá hiệu suất trong quá trình huấn luyện, tối ưu các siêu tham số và kiểm soát overfitting;
- **Tập kiểm thử (test) 10%):** Dùng để đánh giá hiệu năng cuối cùng của mô hình trên dữ liệu chưa từng xuất hiện trong tập huấn luyện.

Việc nhóm dữ liệu theo hình ảnh và phân chia theo tỉ lệ này đảm bảo rằng mô hình VQA được đánh giá một cách chính xác và khách quan, phản ánh đúng khả năng tổng quát hóa trên các di tích mới.

2.2.3. Cài đặt huấn luyện Mô hình Modular (ViT + PhoBERT + ViT5)

Hệ thống được triển khai trên nền tảng Kaggle với GPU Tesla P100 (bộ nhớ VRAM 16GB). Do quy mô dữ liệu hạn chế (5.631 cặp câu hỏi-câu trả lời), nghiên cứu áp dụng phương pháp tinh chỉnh hiệu quả tham số (Parameter-Efficient Fine-Tuning - PEFT) thông qua kỹ thuật LoRA. Cách tiếp cận này chỉ huấn luyện các bộ điều hợp LoRA, mô-đun kết hợp và các lớp chiếu trong khi vẫn giữ nguyên trọng số của ba mô hình nền đã được tiền huấn luyện (ViT, PhoBERT, ViT5).

Bảng 2: Cấu trúc tham số và tỷ lệ tham số tinh chỉnh của mô hình Modular VLM

Thành phần	**Tổng tham số**	**Tham số huấn luyện**	**Tỷ lệ**
ViT base	~86M	295K (LoRA)	0.34%
PhoBERT base	~135M	295K (LoRA)	0.22%
ViT5 base	~220M	590K (LoRA)	0.27%
Mô-đun kết hợp	2.4M	2.4M	100%
Projection	590K	590K	100%
Tổng cộng	***~444M***	***~3.6M***	***0.81%***

Số lượng tham số huấn luyện chỉ chiếm 0.81% tổng số tham số, giúp giảm nhu cầu bộ nhớ GPU từ khoảng 40GB xuống 6-8GB.

2.2.3.1. Chuẩn hóa dữ liệu đầu vào

(1) Xử lý hình ảnh:

Ảnh đầu vào được thay đổi kích thước về 224×224 điểm ảnh và chuẩn hóa theo chuẩn ImageNet với giá trị trung bình là (0,485; 0,456; 0,406) và độ lệch chuẩn là (0,229; 0,224; 0,225). Đầu ra có dạng [B, 3, 224, 224], trong đó B là kích thước batch. Ảnh sau đó được chia thành 196 mảnh theo lưới 14×14, mỗi mảnh có kích thước 16×16 điểm ảnh. ViT xử lý các mảnh này cùng với một token đặc biệt [CLS], tạo ra đầu ra có dạng [B, 197, 768] với kích thước ẩn là 768. Các đặc trưng thị giác cuối cùng được trích xuất bằng cách loại bỏ token [CLS], còn lại [B, 196, 768].

(2) Xử lý câu hỏi tiếng Việt:

Câu hỏi tiếng Việt đầu tiên được phân đoạn từ bằng thư viện PyVi để xử lý đúng các từ nhiều âm tiết. Ví dụ, câu "Bức ảnh này được chụp ở đâu?" được chuyển thành "Bức_ảnh này được chụp ở đâu ?". Chuỗi đã phân đoạn sau đó được token hóa bởi bộ token hóa PhoBERT với độ dài tối đa là 128. Đầu ra bao gồm input_ids và

attention_mask, cả hai đều có dạng [B, 128]. Bộ mã hóa PhoBERT xử lý các token này và tạo ra đặc trưng câu hỏi có dạng [B, 128, 768].

(3) Xử lý câu trả lời:

Câu trả lời được token hóa bằng bộ token hóa ViT5 với độ dài tối đa là 128, tạo ra tensor có dạng [B, 128]. Trong quá trình huấn luyện, các token đệm (có giá trị pad_token_id) được thay bằng -100 để PyTorch bỏ qua khi tính hàm mất mát cross-entropy.

2.2.3.2. Cấu hình LoRA cho các mô-đun

Bảng 3: Cấu hình LoRA áp dụng cho các mô-đun trong mô hình Modular VLM

Model	Rank (r)	Alpha (α)	Dropout	Target Modules	Số adapters
ViT	16	32	0.15	query, key, value	36 (3×12 layers)
PhoBERT	16	32	0.15	query, key, value	36 (3×12 layers)
ViT5	16	32	0.15	q, k, v	72 (3×24 layers)

LoRA hoạt động theo công thức: $h = W_0x + (A \times B)x \times (\alpha/r)$, trong đó W_0 là trọng số gốc được giữ nguyên (frozen), A và B là các ma trận LoRA có thể huấn luyện, với A thuộc $R^{d\times r}$và B thuộc $R^{d\times r}$, còn α/r là hệ số tỷ lệ (scaling). Với cấu hình rank bằng 16 và alpha bằng 32, hệ số tỷ lệ là 2.0.

Thống kê chi tiết cho thấy ViT LoRA có 295,000 tham số (12 layers nhân 3 mô-đun nhân rank 16), PhoBERT LoRA cũng có 295,000 tham số tương tự, ViT5 LoRA có 590,000 tham số (do có 24 layers gồm cả encoder và decoder). Cộng với Fusion Module 2,400,000 tham số và Projection layers 590,000 tham số, tổng số tham số có thể huấn luyện là 3.580.000, chiếm khoảng 0,81% tổng số tham số.

2.2.3.1. Chiến lược tối ưu hóa phân tầng

Bảng 4: Chiến lược thiết lập learning rate cho các thành phần của mô hình Modular VLM

Thành phần	LR gốc	Hệ số nhân	LR cuối	Lý do
ViT LoRA	1e-4	1.0×	1e-4	Tốc độ chuẩn
PhoBERT LoRA	1e-4	0.5×	5e-5	Bảo toàn tri thức tiền huấn luyện
ViT5 LoRA	1e-4	0.3×	3e-5	Tránh quên thảm họa
Fusion Module	1e-4	5.0×	5e-4	Mô-đun mới, học nhanh

Hệ thống áp dụng chiến lược tối ưu hóa phần tầng cho từng thành phần để tối ưu quá trình huấn luyện. Learning rate gốc là 1e-4, sau đó được điều chỉnh bởi các hệ số nhân khác nhau. ViT LoRA giữ nguyên learning rate 1e-4. PhoBERT LoRA sử dụng hệ số nhân 0.5 để có learning rate 5e-5, nhằm bảo toàn tri thức tiền huấn luyện. ViT5 LoRA có hệ số nhân thấp nhất là 0.3 cho learning rate 3e-5, vì decoder cực kỳ nhạy cảm và cần tránh hiện tượng quên thảm họa (catastrophic forgetting). Ngược lại, Fusion Module là thành phần hoàn toàn mới chưa được tiền huấn luyện nên sử dụng hệ số nhân 5.0 cho learning rate 5e-4 để học nhanh các quan hệ đa phương thức.

Bộ tối ưu sử dụng là AdamW với hệ số suy giảm trọng số (weight decay) 0.01, betas (0.9, 0.999), epsilon 1e-8, và max gradient tối đa 1.0 cho việc cắt gradient. Learning rate schedule theo kiểu linear warmup trong 5% tổng số steps đầu tiên, sau đó decay tuyến tính từ peak về 0.

2.2.3.2. Thiết kế Fusion Module với cơ chế Gating

Fusion Module nhận ba đầu vào: visual_features có dạng [B, 196, 768] từ ViT, question_features có dạng [B, seq_len, 768] từ PhoBERT (với seq_len không quá 128), và question_mask có dạng [B, seq_len].

Quá trình xử lý diễn ra qua ba bước chính.

Bước 1 là cross-attention hai chiều: visual features chú ý question features tạo ra tensor có dạng [B, seq_len, 768], đồng thời question features chú ý visual features tạo ra tensor có dạng [B, 196, 768].

Bước 2 áp dụng cơ chế gating để điều chỉnh trọng số giữa hai nguồn thông tin. Cụ thể, attended_visual được pooling theo chiều thứ nhất (dim=1) bằng phép lấy trung bình để tạo vis_pooled có có dạng [B, 768]. Tương tự, attended_question cũng được pooling tạo q_pooled có có dạng [B, 768]. Hai vectors này được concatenate thành gate_input có shape [B, 1536], sau đó đưa qua một mạng fully-connected và hàm sigmoid để tạo gate_weight có shape [B, 1]. Gate weight này được sử dụng để nhân với attended_visual, trong khi phần bù (1 trừ gate_weight) được nhân với attended_question.

Bước 3 là concatenation và projection: attended_visual và attended_question sau khi đã được điều chỉnh bởi gate được concatenate theo chiều thứ nhất, tạo ra fused tensor có có dạng [B, (196 + seq_len), 768]. Tensor này cuối cùng được chiếu về không gian hidden size của ViT5 decoder.

Cấu hình của module bao gồm 8 attention heads, dropout 0.15, và gate network gồm các lớp Linear(1536→768), ReLU, Dropout, Linear(768→1), và Sigmoid.

2.2.3.3. Tham số huấn luyện

Bảng 5: Cấu hình huấn luyện mô hình Modular VLM

Tham số	**Giá trị**	**Ghi chú**
Batch size	4	Batch vật lý
Gradient accumulation	4	Effective batch = 16
Effective batch size	16	4 × 4
Số epoch tối đa	20	Với early stopping
Số epoch thực tế	13	Dừng sớm do early stopping
Learning rate (base)	1e-4	Cho ViT LoRA
Tỷ lệ warmup	0.05	5% bước đầu
Weight decay	0.01	Điều chuẩn L2
Chuẩn gradient tối đa	1.0	Cắt gradient
Mixed precision	FP16	Giảm ~40–50% bộ nhớ
Gradient checkpointing	Enabled	Tiết kiệm bộ nhớ
Số worker	2	Số tiến trình tải dữ liệu

Batch size vật lý được đặt ở mức 4, kết hợp với gradient accumulation qua 4 bước tạo ra effective batch size là 16. Mô hình được cấu hình huấn luyện tối đa 20 epoch nhưng thực tế dừng lại sau epoch thứ 13 do early stopping. Learning rate cơ sở cho ViT LoRA là 1e-4, với tỷ lệ warmup 5% trong các bước đầu tiên. Weight decay 0,01 được áp dụng cho chuẩn hóa L2, và chuẩn gradient tối đa 1,0 được sử dụng cho việc cắt gradient.

Các kỹ thuật tối ưu bộ nhớ bao gồm mixed precision FP16 giúp giảm 40-50% dung lượng bộ nhớ, gradient checkpointing giảm thêm khoảng 30% bộ nhớ activation, và gradient accumulation cho phép mô phỏng batch lớn với bộ nhớ nhỏ. Tổng dung lượng bộ nhớ sử dụng trên GPU là 6-8GB trên Tesla P100 16GB.

2.2.3.4. Cơ chế dừng sớm dựa trên điểm tổng hợp

Thay vì chỉ dựa vào một chỉ số đánh giá đơn lẻ, hệ thống sử dụng điểm tổng hợp (composite score), kết hợp năm chỉ số với trọng số khác nhau theo công thức:

Composite Score = 0.30 × BLEU + 0.30 × BERTScore-F1 + 0.20 × CIDEr + 0.15 × ROUGE-L + 0.05 × Exact Match

Bảng 6: Cấu trúc trọng số của các chỉ số đánh giá dùng để đánh giá mô hình

Chỉ số đánh giá	Weight	Ý nghĩa
BLEU	30%	Độ chính xác từ vựng (n-gram)
BERTScore-F1	30%	Độ tương đồng ngữ nghĩa sâu
CIDEr	20%	Dựa trên sự đồng thuận (TF-IDF)
ROUGE-L	15%	Chuỗi con chung dài nhất
Exact Match	5%	Khớp hoàn toàn

Phân bổ trọng số này đảm bảo cân bằng giữa độ chính xác từ vựng (BLEU chiếm 30%, Exact Match chiếm 5%) và độ tương đồng ngữ nghĩa (BERTScore-F1 chiếm 30%, CIDEr chiếm 20%). ROUGE-L với 15% đo lường chuỗi con chung dài nhất để đánh giá cấu trúc câu.

Early stopping được cấu hình với patience 3 epoch, nghĩa là nếu các điểm tổng hợp trên tập kiểm tra không cải thiện sau 3 epoch liên tiếp thì quá trình huấn luyện sẽ dừng lại. Chế độ được thiết lập là tối đa hóa với min delta 0,001. Trong thực tế, mô hình tốt nhất được lưu lại ở epoch thứ 7 và quá trình huấn luyện dừng lại sau epoch thứ 13 do không có cải thiện thêm trong 6 epoch tiếp theo.

2.2.3.5. Cấu hình sinh câu trả lời

Bảng 7: Cấu hình giải mã (decoding configuration) cho mô hình sinh câu trả lời

Tham số	Giá trị	Mục đích
Strategy	Beam search	Tìm kiếm tốt hơn greedy
num_beams	3	Số lượng beams
max_length	128	Độ dài tối đa
min_length	3	Tránh câu quá ngắn
no_repeat_ngram_size	2	Không lặp bigram
length_penalty	0.8	Khuyến khích câu ngắn
repetition_penalty	1.2	Phạt từ lặp lại
early_stopping	True	Dừng khi đủ beams
do_sample	False	Xác định

Quá trình sinh câu trả lời sử dụng beam search với 3 beam thay vì giải mã tham lam (greedy decoding) để có chất lượng tốt hơn. Độ dài đầu ra được giới hạn trong khoảng từ 3 đến 128 token. Tham số no_repeat_ngram_size bằng 2 ngăn chặn việc lặp lại bất kỳ bigram nào trong câu sinh ra. Length penalty 0,8 khuyến khích mô hình sinh câu ngắn gọn hơn, trong khi repetition penalty 1,2 phạt các từ lặp lại để tăng tính đa dạng. Early stopping trong beam search được kích hoạt để dừng ngay khi tất cả beam đều đạt token kết thúc chuỗi. Do_sample được đặt False để đảm bảo quá trình sinh hoàn toàn xác định.

Beam search strategy duy trì 3 candidate sequences song song, tại mỗi bước chọn token có log-probability cao nhất, loại bỏ các beams có cumulative probability thấp, và dừng lại khi tất cả beams đạt EOS token.

2.2.4. Cài đặt giải pháp mô hình Qwen2-VL

Mô hình Qwen2-VL-7B-Instruct với 7 tỷ tham số được huấn luyện theo kiến trúc Unified. Do quy mô lớn và giới hạn GPU (Tesla P100 16GB), nghiên cứu kết hợp ba kỹ thuật tối ưu:

Bảng 8: Hiệu quả các kỹ thuật tối ưu bộ nhớ và tham số trong quá trình huấn luyện

Kỹ thuật	**Hiệu quả**	**Kết quả**
4-bit Quantization	Bộ nhớ giảm 75%	28GB xuống 7GB
LoRA (r=32)	Tham số huấn luyện giảm 99,87%	7B xuống 9M
Gradient Checkpointing	Bộ nhớ activation giảm 30%	Chạy vừa trong 16GB

Bảng 9: Phân bố tham số và tỷ lệ tham số có thể huấn luyện trong mô hình Unified VLM

Thành phần	**Params**	**Trainable**	**Tỷ lệ**
Base model (quantized)	7B	0	0%
LoRA adapters	9M	9M	100%
Tổng cộng	***7.009B***	***9M***	***0.13%***

Kỹ thuật lượng tử hóa 4-bit giảm dung lượng bộ nhớ 75% từ 28GB xuống 7GB. LoRA với hạng 32 chỉ huấn luyện 9 triệu tham số, giảm 99,87% so với toàn bộ 7 tỷ tham số của base model. Gradient checkpointing giảm thêm 30% bộ nhớ activation, giúp toàn bộ mô hình vừa khít trong GPU 16GB.

2.2.4.1. Mẫu hội thoại và xử lý token hình ảnh

Qwen2-VL sử dụng mẫu hội thoại đặc biệt để định dạng đầu vào. Mỗi mẫu dữ liệu được cấu trúc thành một cuộc hội thoại với hai vai trò: user và assistant. Phần user bao gồm hình ảnh (dưới dạng đối tượng PIL Image) kết hợp với câu hỏi văn bản. Phần assistant chỉ chứa câu trả lời văn bản. Ví dụ, với câu hỏi "Bức ảnh này được chụp ở đâu?", template user sẽ có content là một dictionary với type "image" chứa đối tượng PIL Image và một dictionary với type "text" chứa câu hỏi. Template assistant có content là dictionary với type "text" chứa câu trả lời "Bức ảnh được ghi lại tại Căn cứ Tỉnh ủy Cần Thơ." (*Hình 16*).

```
messages = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": <PIL.Image>},
            {"type": "text", "text": "Bức ảnh này được chụp ở đâu?"}
        ]
    },
    {
        "role": "assistant",
        "content": [
            {"type": "text", "text": "Bức ảnh được chụp tại Căn cứ rừng U Minh."}
        ]
    }
]
```

Hình 16: Template sử dụng cho mô hình Qwen2-VL

Sau khi áp dụng chat template, chuỗi được chuyển đổi thành định dạng đặc biệt với các token đặc biệt. Chuỗi bắt đầu bằng token im_start theo sau là "system", system prompt "You are a helpful assistant", và token im_end. Tiếp theo là token im_start, "user", các visual token được bao bọc bởi vision_start và vision_end (chứa các token image_pad), câu hỏi văn bản, và token im_end. Cuối cùng là token im_start, "assistant", câu trả lời, và token im_end (*Hình 17*).

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|image_pad|>...<|vision_end|>
Bức ảnh này được chụp ở đâu?<|im_end|>
<|im_start|>assistant
Bức ảnh được ghi lại tại Căn cứ Tỉnh ủy Cần Thơ..<|im_end|>
```

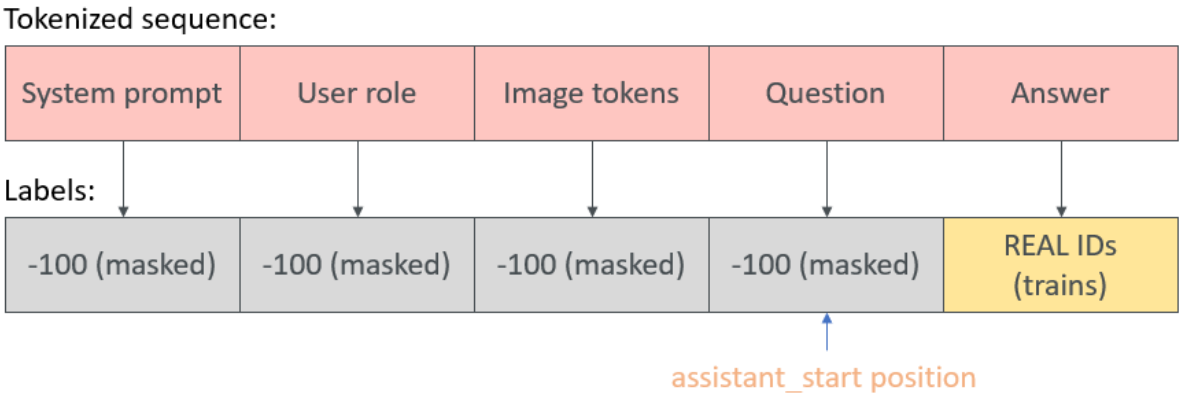
Hình 17: Định dạng prompt multimodal của mô hình Qwen2-VL

Hình ảnh được xử lý với độ phân giải động, giữ nguyên tỷ lệ khung hình trong khi thay đổi kích thước về tối đa 336×336 điểm ảnh. Ảnh có kênh alpha (RGBA, P, LA) được chuyển sang RGB. Vision encoder dựa trên Vision Transformer trích xuất đặc trưng và chuyển thành chuỗi các visual token có dạng [B, num_visual_tokens, hidden_size], trong đó num_visual_tokens khoảng 200-300 tùy thuộc độ phân giải và hidden_size là 3584 cho Qwen2-VL. Các visual token này được nối với text token để tạo thành chuỗi thống nhất có dạng [B, total_seq_len, 3584], với total_seq_len bằng tổng của num_visual_tokens và num_text_tokens, thường khoảng 400-500 token.

2.2.4.2. Chiến lược Label Masking

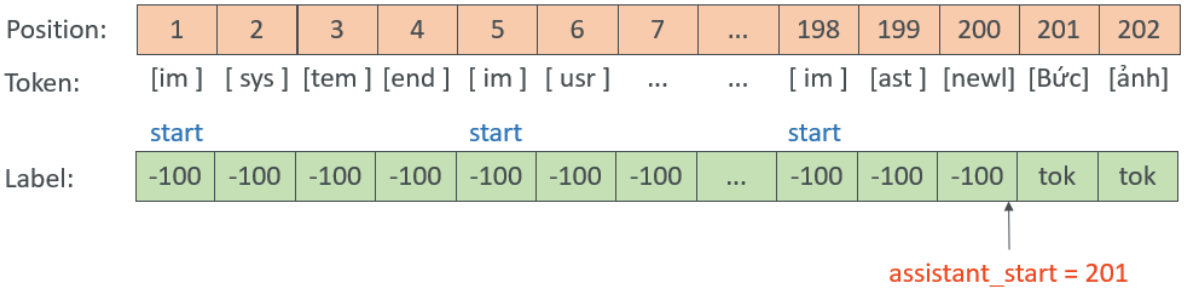
Chiến lược label masking là một kỹ thuật quan trọng nhằm đảm bảo mô hình chỉ học cách sinh câu trả lời mà không ghi nhớ prompt template. Quá trình bắt đầu bằng việc token hóa toàn bộ cuộc hội thoại thành một chuỗi input_ids duy nhất. Chuỗi này bao gồm các phần theo thứ tự: system prompt, user role marker, image token, câu hỏi, và câu trả lời.

Hệ thống cần xác định chính xác vị trí bắt đầu của phần câu trả lời (assistant_start position). Thuật toán tìm kiếm duyệt qua chuỗi input_ids từ trái sang phải để phát hiện mẫu cụ thể. Đầu tiên, tìm token im_start_id đánh dấu vai trò. Sau đó kiểm tra các token tiếp theo có đánh vần "assistant" hay không. Nếu khớp, thuật toán bỏ qua token newline_id (xuống dòng), và vị trí ngay sau newline chính là assistant_start (*Hình 18*).



Hình 18: Cách Qwen2-VL mã hóa token và áp dụng label masking trong huấn luyện

Ví dụ minh họa (*Hình 19*): tại vị trí 198 là token im_start, vị trí 199 là token "ast" (assistant), vị trí 200 là token newline, thì assistant_start sẽ là vị trí 201 nơi bắt đầu token "Bức". Tất cả các vị trí từ 0 đến 200 (bao gồm system prompt, user role, image token, câu hỏi, và assistant role markers) được gán nhãn -100. Chỉ có các token từ vị trí 201 trở đi (phần câu trả lời) mới giữ nguyên token ID làm nhãn để tính loss..



Hình 19: Minh họa tokenization và label masking, kèm vị trí assistant_start trong chuỗi đầu vào

Hệ thống xác minh bằng cách đếm số token có nhãn khác -100. Nếu bằng 0, toàn bộ mẫu bị bỏ qua bằng cách gán tất cả nhãn thành -100. Nếu lớn hơn 0, mẫu được coi là hợp lệ và sử dụng cho huấn luyện. Thống kê trên một batch điển hình cho thấy: với batch size 1 và tổng số 512 token, có 289 prompt token bị che (56,4%), 223 answer token có thể huấn luyện (43,6%), và 0 padding token.

2.2.4.3. 4-bit Quantization (QLoRA)

Bảng 10: Cấu hình 4-bit quantization (BitsAndBytes) cho mô hình

Tham số	**Giá trị**	**Giải thích**
load_in_4bit	True	Lượng tử hóa trọng số với 4-bit
bnb_4bit_compute_dtype	bfloat16	Forward/backward ở bfloat16
bnb_4bit_use_double_quant	True	Lượng tử hóa cả hằng số
bnb_4bit_quant_type	nf4	Phân phối NormalFloat4

Bảng 11: Mức tiêu thụ bộ nhớ của mô hình dưới các thiết lập precision khác nhau

Precision	**Memory**	**Giảm so với FP32**
FP32 (full)	~28GB	0%
FP16 (half)	~14GB	50%
Int8 (8-bit)	~7GB	75%
NF4 (4-bit)	~3.5GB	87.5%
+ LoRA adapters	+4GB	-
Tổng cộng	***~7-8GB***	***71-75%***

Qwen2-VL áp dụng lượng tử hóa (quantization) 4-bit thông qua thư viện BitsAndBytes. Cấu hình load_in_4bit được kích hoạt để lượng tử hóa tất cả trọng số xuống 4-bit, trong khi bnb_4bit_compute_dtype được đặt là bfloat16 để đảm bảo các phép tính lan truyền xuôi và lan truyền ngược vẫn diễn ra ở độ chính xác cao hơn. Tham số bnb_4bit_use_double_quant được bật, nghĩa là ngay cả các hằng số lượng tử hóa cũng được lượng tử hóa thêm lần nữa. Loại lượng tử hóa là NormalFloat4 (nf4), một phương pháp được thiết kế tối ưu cho phân phối trọng số của các mô hình tiền huấn luyện.

Quy trình lượng tử hóa diễn ra như sau: trọng số gốc ở FP32 thuộc miền từ âm vô cùng đến dương vô cùng được chuẩn hóa về khoảng -1 đến 1 bằng cách chia cho giá trị tuyệt đối lớn nhất. Trọng số đã chuẩn hóa được lượng tử hóa thành NF4 với 16 giá trị rời rạc được tối ưu cho phân phối chuẩn. Kết quả được lưu dưới dạng W_quant (4-bit) cộng với hệ số tỷ lệ (16-bit). Khi cần tính toán, trọng số được giải lượng tử hóa bằng cách nhân W_quant với hệ số tỷ lệ, sau đó phép tính diễn ra ở bfloat16 với activation bằng W_deq nhân input.

So sánh dung lượng bộ nhớ cho thấy FP32 độ chính xác đầy đủ tiêu tốn khoảng 28GB, FP16 độ chính xác nửa giảm 50% xuống 14GB, Int8 lượng tử hóa 8-bit giảm 75% xuống 7GB, và NF4 lượng tử hóa 4-bit giảm 87,5% xuống chỉ 3,5GB. Cộng thêm khoảng 4GB cho LoRA adapters, tổng dung lượng bộ nhớ sử dụng khoảng 7-8GB, giảm 71-75% so với FP32 và vừa khít với GPU 16GB.

2.2.4.4. Cấu hình LoRA

Bảng 12: Cấu hình LoRA của Unified VLM (Qwen2-VL) và so sánh với mô hình Modular

Tham số	**Giá trị**	**So sánh Modular**
Rank (r)	32	Gấp đôi (16)
Alpha (α)	64	Gấp đôi (32)
Scaling (α/r)	2.0	Giống nhau
Dropout	0.1	Thấp hơn (0.15)
Loại tác vụ	CAUSAL_LM	Thấp hơn (0.15)

LoRA được áp dụng với hạng 32 và alpha 64, cao hơn gấp đôi so với mô hình Modular (hạng 16, alpha 32). Lý do là Qwen2-VL có 7 tỷ tham số, lớn hơn nhiều so với tổng ba mô hình trong Modular (khoảng 400 triệu), do đó cần dung lượng cao hơn để thích ứng hiệu quả. Hệ số tỷ lệ α/r vẫn giữ nguyên ở mức 2,0. Dropout được giảm xuống 0,1 thay vì 0,15 vì base model đã rất lớn. Loại tác vụ là CAUSAL_LM thay vì SEQ2SEQ như ViT5.

Target modules bao gồm tất cả các phép chiếu attention (q_proj, k_proj, v_proj, o_proj) trong 28 lớp Transformer, tạo ra 112 adapters. Các phép chiếu MLP (gate_proj, up_proj, down_proj) cũng được gắn LoRA, thêm 84 adapters. Nếu vision encoder cũng có LoRA, các modules qkv, proj, fc1, fc2 sẽ được thêm adapters. Tổng cộng có khoảng 200 LoRA adapter modules trên toàn bộ mô hình.

Chiến lược đóng băng được áp dụng: vision encoder không có LoRA được đóng băng hoàn toàn, chỉ có text encoder/decoder với LoRA adapters mới được huấn luyện (backbone vẫn bị đóng băng). Tổng số tham số có thể huấn luyện là khoảng 9 triệu. Gradient checkpointing được kích hoạt với đánh đổi giảm 30% bộ nhớ nhưng tăng 20% thời gian tính toán. Các điểm kiểm tra được đặt mỗi 4 lớp, các activation bị thiếu sẽ được tính toán lại trong quá trình lan truyền ngược. Tham số use_cache phải được tắt vì không tương thích với gradient checkpointing.

2.2.4.5. Cấu hình và Chiến lược Huấn luyện

Bảng 13: Cấu hình huấn luyện mô hình Unified VLM (Qwen2-VL) và so sánh với mô hình Modular VLM

Tham số	**Modular**	**Unified**	**Tỷ lệ khác biệt**
Batch size	4	1	0.25×
Gradient accumulation	4	32	8×
Effective batch	16	32	2×
Learning rate	1e-4	5e-6	0.05× (thấp hơn 20×)
LR scheduler	Linear	Cosine	
Tỷ lệ warmup	5%	5%	Giống nhau
Weight decay	0.01	0.01	Giống nhau
Số epoch tối đa	20	3	0.15×
Số epoch thực tế	13	3	0.23×
Early stop patience	3	2	0.67×
Độ dài chuỗi tối đa	128	640	5×
Thời gian huấn luyện/epoch	~25 min	~45 min	1.8×

Do base model rất lớn, batch size vật lý chỉ có thể đặt là 1. Để bù đắp, gradient accumulation được tăng lên 32 bước, tạo ra effective batch size là 32, lớn hơn Modular có effective batch 16. Learning rate được đặt ở mức rất thấp 5e-6, thấp hơn 20 lần so

với Modular (1e-4). Lý do là mô hình 7B tham số cực kỳ nhạy cảm với learning rate cao. Learning rate quá cao (lớn hơn 1e-5) sẽ làm loss diverge và tạo ra gradient NaN. Learning rate tối ưu 5e-6 nằm trong khoảng được khuyến nghị bởi bài báo QLoRA (từ 1e-5 đến 5e-6), đủ nhỏ để ổn định nhưng đủ lớn để hội tụ trong 3 epoch.

Learning rate schedule sử dụng cosine annealing thay vì suy giảm tuyến tính. Sau giai đoạn warmup 5% đầu tiên, learning rate giảm theo đường cong cosin từ đỉnh 5e-6 xuống 0 ở cuối quá trình huấn luyện, tạo ra sự giảm mượt mà hơn. Tỷ lệ warmup và weight decay giữ nguyên như Modular ở mức 5% và 0,01.

Số epoch chỉ là 3, ngắn hơn rất nhiều so với Modular (20 epoch, dừng ở 13), phản ánh việc mô hình lớn hội tụ nhanh hơn nhờ nền tảng tiền huấn luyện mạnh. Patience của early stopping cũng giảm xuống 2 epoch thay vì 3 để phù hợp với tổng số epoch ít hơn. Độ dài chuỗi tối đa được tăng lên 640 token (so với 128 của Modular) để có thể xử lý visual token (200-300 token) cộng với câu hỏi và câu trả lời dài hơn. Thời gian huấn luyện mỗi epoch khoảng 45 phút, chậm hơn 1,8 lần so với Modular (25 phút).

2.2.4.6. Đánh giá và Chiến lược Sinh câu trả lời

Bảng 14: Tham số giải mã của mô hình Unified VLM (Qwen2-VL) và so sánh với kiến trúc Modular VLM

Tham số	**Modular**	**Unified**	**Lý do khác**
Chiến lược	Beam search	Greedy	Bộ nhớ & tốc độ
num_beams	3	1 (greedy)	-
max_new_tokens	128	256	dài hơn 2×
min_length	3	-	-
temperature	-	None (disabled)	Xác định
top_p	-	None	Xác định
top_k	-	None	Xác định
repetition_penalty	1.2	1.2	Giống nhau
do_sample	False	False	Giống nhau

Khác với Modular sử dụng beam search với 3 beams, Qwen2-VL chỉ dùng greedy decoding (chọn token có xác suất cao nhất tại mỗi bước). Lý do là mô hình 7B đã sinh rất tốt với greedy, trong khi beam search tốn gấp 3 lần bộ nhớ (phải duy trì 3 chuỗi song song) và chậm hơn đáng kể. Beam search với 3 beam cần bộ nhớ gấp 3 lần để duy trì 3 chuỗi, chậm hơn 3 lần về thời gian, nhưng chỉ cải thiện chất lượng khoảng 1-2%

(marginal gain). Ngược lại, greedy decoding chỉ cần bộ nhớ 1× cho chuỗi đơn, nhanh nhất (1× time), và chất lượng đã đủ tốt với LLM 7B.

Độ dài đầu ra tối đa được tăng lên 256 token (gấp đôi Modular) để cho phép câu trả lời chi tiết hơn. Temperature, top_p, và top_k đều được tắt (đặt None) để quá trình sinh hoàn toàn xác định. Chỉ có repetition penalty ở mức 1,2 (giống Modular) được giữ lại để tránh lặp từ. Do_sample được đặt False để đảm bảo tính xác định.

Trong quá trình đánh giá trên tập kiểm th, có một chi tiết kỹ thuật quan trọng: batch size được tăng lên 8 (thay vì 1 trong huấn luyện) vì không cần lan truyền ngược nên tiết kiệm bộ nhớ, và phía padding phải được chuyển từ phải sang trái. Trong huấn luyện với padding bên phải, các chuỗi có dạng [prompt_1][answer_1][PAD][PAD] và [prompt_2][answer_2][answer_2][PAD], khiến quá trình sinh bắt đầu ở các vị trí khác nhau (không căn chỉnh). Trong sinh với padding bên trái, các chuỗi có dạng [PAD][PAD][prompt_1] và [PAD][prompt_2], khiến quá trình sinh bắt đầu cùng một vị trí (đã căn chỉnh). Padding bên trái đảm bảo tất cả chuỗi trong batch kết thúc prompt ở cùng vị trí, sau đó quá trình sinh bắt đầu đồng thời.

Khi giải mã đầu ra, hệ thống chỉ lấy phần sau prompt (bỏ qua input_ids) để chỉ giữ lại câu trả lời được sinh. Chuỗi đầu ra có dạng [input_ids | generated_ids | pad_tokens], câu trả lời cuối cùng được tạo bằng cách giải mã từ vị trí len(input_ids) trở đi với skip_special_tokens=True. Bộ nhớ đệm được xóa định kỳ sau mỗi 10 batch để tránh hết bộ nhớ khi chạy trên toàn bộ tập kiểm thử.

CHƯƠNG 3. THỰC NGHIỆM VÀ ĐÁNH GIÁ TRỢ LÝ ẢO DI SẢN VĂN HÓA CÁCH MẠNG DỰA TRÊN HÌNH ẢNH

3.1. Giao diện sản phẩm

3.1.1. Trang chủ

Hình 20: Giao diện trang chủ

Trang chủ (*Hình 20*) được xây dựng dưới dạng giao diện tĩnh, đóng vai trò điểm truy cập ban đầu vào hệ thống. Bố cục bao gồm thanh menu bên trái và khu vực nội dung trung tâm dùng để hiển thị các mục thông tin cơ bản như tin tức và sự kiện. Giao diện không chứa các chức năng tương tác phức tạp mà chủ yếu đóng vai trò điều hướng, dẫn người dùng đến các trang chức năng khác trong hệ thống.

3.1.2. Trang thông tin di sản văn hóa

Hình 21: Giao diện trang thông tin di sản văn hóa (1)

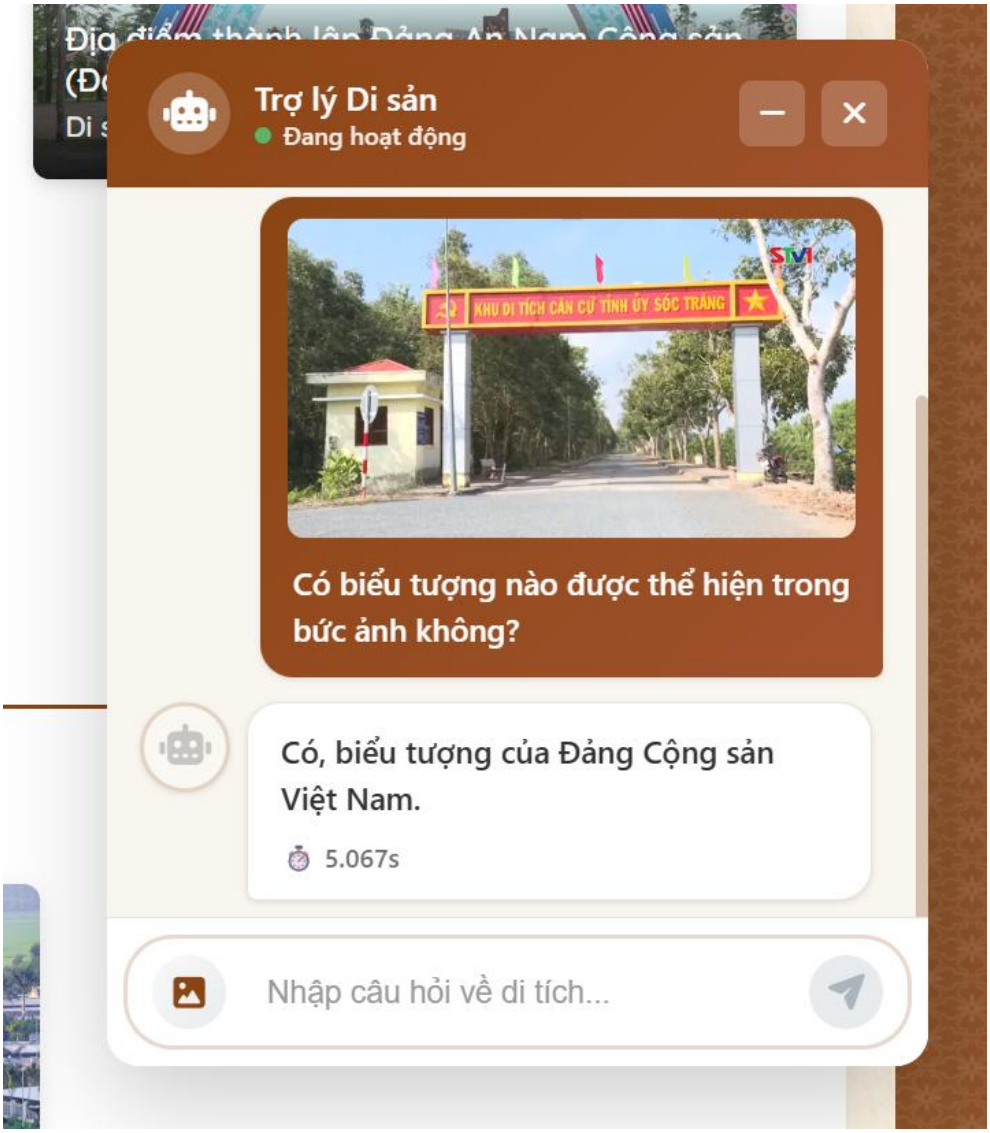
Hình 22: Hình 21: Giao diện trang thông tin di sản văn hóa (2)

Trang Thông tin Di sản Văn hóa (*Hình 21*) được thiết kế theo dạng danh sách thẻ (card), trong đó mỗi thẻ bao gồm hình ảnh và mô tả ngắn. Giao diện này được xây dựng với mục đích mô phỏng và hoàn thiện cấu trúc hệ thống; không phải là trọng tâm của đề tài mà chỉ mang tính minh họa, hỗ trợ phần demo chức năng của Trợ lý Ảo.

3.1.3. Giao diện trợ lý ảo

Hình 23: Giao diện trợ lí ảo được tích hợp vào trang web di sản văn hóa

Giao diện Trợ lý Ảo được bố trí tại góc dưới bên phải màn hình (*Hình 23*), với thiết kế tối giản nhằm đảm bảo dễ sử dụng và thuận tiện trong quá trình tương tác. Giao diện được chia thành ba khu vực chính như sau:



Hình 24: Giao diện trợ lí ảo

(1) Thanh tiêu đề

Thanh tiêu đề hiển thị tên hệ thống "Trợ lý Di sản", kèm biểu tượng chấm xanh cho biết trạng thái hoạt động của chatbot. Người dùng có thể quản lý cửa sổ trò chuyện thông qua hai nút điều khiển:

- **Nút thu nhỏ (-):** Thu gọn cửa sổ nhưng vẫn giữ nguyên toàn bộ nội dung hội thoại, cho phép mở lại và tiếp tục tương tác.
- **Nút đóng (x):** Đóng cửa sổ và đồng thời xóa toàn bộ lịch sử hội thoại trước đó.

Thiết kế này giúp người dùng linh hoạt trong việc bật/tắt giao diện trợ lý mà không ảnh hưởng đến trải nghiệm chung trên trang.

(2) Khu vực hiển thị nội dung hội thoại

Khu vực này trình bày toàn bộ luồng trao đổi giữa người dùng và hệ thống. Các thành phần được hiển thị theo dạng bong bóng hội thoại nhằm đảm bảo tính trực quan và dễ theo dõi:

- Ảnh người dùng tải lên được hiển thị lại trong luồng hội thoại, giúp đối chiếu trực tiếp với câu trả lời của hệ thống.
- Câu hỏi của người dùng được trình bày rõ ràng, được căn chỉnh và tách biệt với phản hồi của chatbot để dễ nhận biết.
- Câu trả lời của AI được hiển thị kèm biểu tượng chatbot và thời gian xử lý (ví dụ: 5.067s), thể hiện tốc độ phản hồi của mô hình.

Bố cục này giúp người dùng theo dõi toàn bộ nội dung tương tác một cách liền mạch và trực quan.

(3) Khu vực tương tác chính (Main Interaction Area)

Khu vực tương tác chính là nơi người dùng thực hiện toàn bộ các thao tác nhập liệu và gửi yêu cầu đến hệ thống. Giao diện được thiết kế trực quan, tập trung vào hai dạng dữ liệu đầu vào: hình ảnh và văn bản. Các thành phần chính bao gồm:

- **Tải ảnh di tích:** Người dùng chọn ảnh từ thiết bị cá nhân. Ngay sau khi tải lên, ảnh được hiển thị trong khung xem trước (preview) để kiểm tra lại nội dung trước khi gửi yêu cầu.
- **Nhập câu hỏi:** Ô nhập văn bản cho phép người dùng đặt câu hỏi liên quan đến ảnh đã tải lên. Hệ thống hỗ trợ nhập tiếng Việt, phù hợp với ngữ cảnh các câu hỏi về di sản văn hóa.
- **Nút gửi yêu cầu:** Nút gửi được kích hoạt khi cả ảnh và câu hỏi đã được cung cấp hợp lệ. Cơ chế này giúp đảm bảo quy trình tương tác diễn ra đúng thứ tự và tránh tình trạng gửi yêu cầu rỗng.
- **Xử lý và cập nhật lên luồng hội thoại:** Sau khi người dùng gửi yêu cầu, hình ảnh, câu hỏi và phản hồi của mô hình được đưa vào khu vực hiển thị nội dung hội thoại ở phía trên. Điều này tạo thành một luồng trao đổi liên tục và dễ theo dõi.

Khu vực tương tác chính đóng vai trò là điểm khởi phát toàn bộ quá trình làm việc giữa người dùng và hệ thống, đảm bảo thao tác rõ ràng, mạch lạc và phù hợp với quy trình xử lý đa phương thức của mô hình VQA.

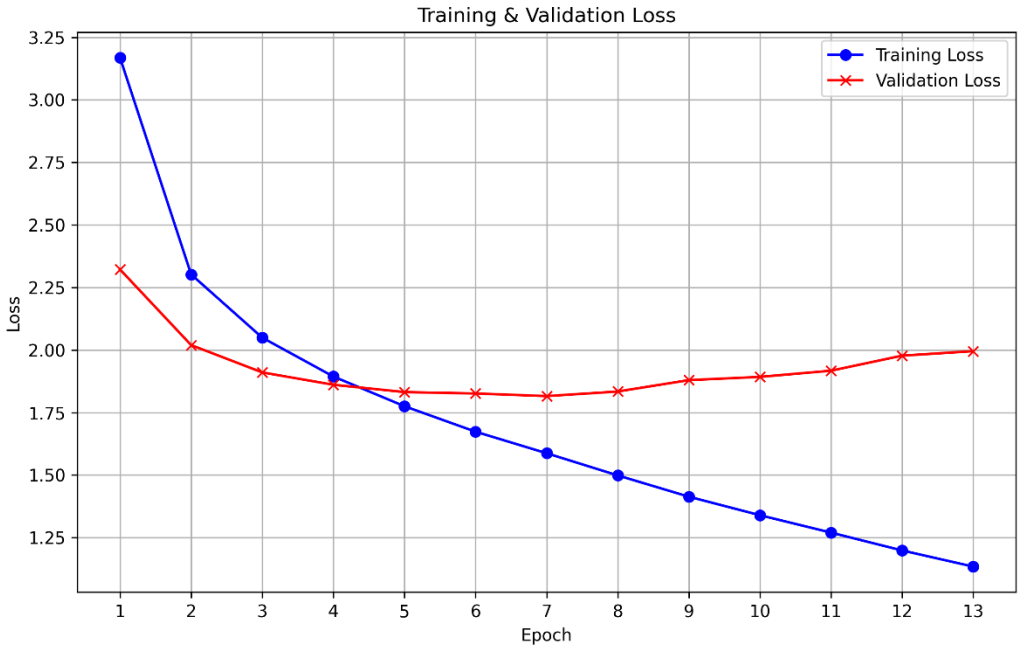
3.2. Kết quả thực nghiệm

Quá trình huấn luyện và đánh giá hai kiến trúc VQA – Modular (ViT + PhoBERT + ViT5) và Unified (Qwen2-VL) – được thực hiện trên cùng tập dữ liệu kiểm thử nhằm đảm bảo tính nhất quán trong phân tích. Các thí nghiệm tập trung đánh giá khả năng mô

hình học biểu diễn hình ảnh và văn bản, cũng như hiệu quả sinh câu trả lời theo cả góc độ từ vựng và ngữ nghĩa.

3.2.1. Kết quả huấn luyện mô hình ViT, PhoBERT và ViT5

Hình 25 minh họa diễn biến giá trị training loss và validation loss của kiến trúc Modular qua các epoch.



Hình 25: Biểu đồ loss mô hình ViT, PhoBERT và ViT5

Training loss giảm đều từ khoảng 3.18 tại epoch 1 xuống 1.13 tại epoch 13, phản ánh quá trình tối ưu hội tụ ổn định và khả năng mô hình học được các đặc trưng quan trọng của dữ liệu huấn luyện.

Validation loss giảm từ mức ban đầu ~2.33 xuống khoảng 1.80 tại các epoch 5–7 trước khi tăng nhẹ và dao động quanh 1.85–2.00. Sự dao động này xuất phát từ kích thước tập kiểm định nhỏ hơn, khiến giá trị loss nhạy cảm với biến thiên dữ liệu. Mặc dù có dao động, mức chênh lệch giữa training và validation loss vẫn cho thấy mô hình duy trì khả năng tổng quát hóa hợp lý.

Bảng 15 trình bày các chỉ số đánh giá hiệu suất trên tập kiểm thử:

Bảng 15: Kết quả đánh giá hiệu suất mô hình ViT, PhoBERT và ViT5 trên tập kiểm thử

BLUE	ROUGE1	ROUGE2	ROUGEL	BERTSCORE-F1	CIDER	EM
0.11	0.61	0.31	0.44	0.81	0.79	0.00

Phân tích kết quả:

- **BLEU = 0.11** phản ánh mức độ trùng khớp n-gram thấp, phù hợp với đặc thù VQA, nơi câu trả lời đúng có thể biểu đạt đa dạng.
- **ROUGE-1 đạt 0.61**, cho thấy mô hình nắm bắt tốt các từ khóa quan trọng; trong khi **ROUGE-2 và ROUGE-L** thấp hơn cho thấy khả năng ghi nhớ chuỗi từ và cấu trúc câu còn hạn chế.
- **BERTScore-F1 = 0.81 và CIDEr = 0.79** cho thấy sự tương đồng ngữ nghĩa cao giữa câu trả lời sinh ra và câu tham chiếu, chứng minh mô hình hiểu chính xác nội dung câu hỏi và hình ảnh ngay cả khi câu trả lời không trùng khớp hoàn toàn về mặt từ vựng.
- **Exact Match = 0.00** là kết quả phổ biến trong VQA do tính đa dạng trong cách biểu đạt câu trả lời đúng.

Hình 26 mô tả các ví dụ dự đoán thực tế, minh chứng rằng mô hình sinh câu trả lời có tính nhất quán ngữ nghĩa với câu tham chiếu, ngay cả khi khác biệt từ ngữ.

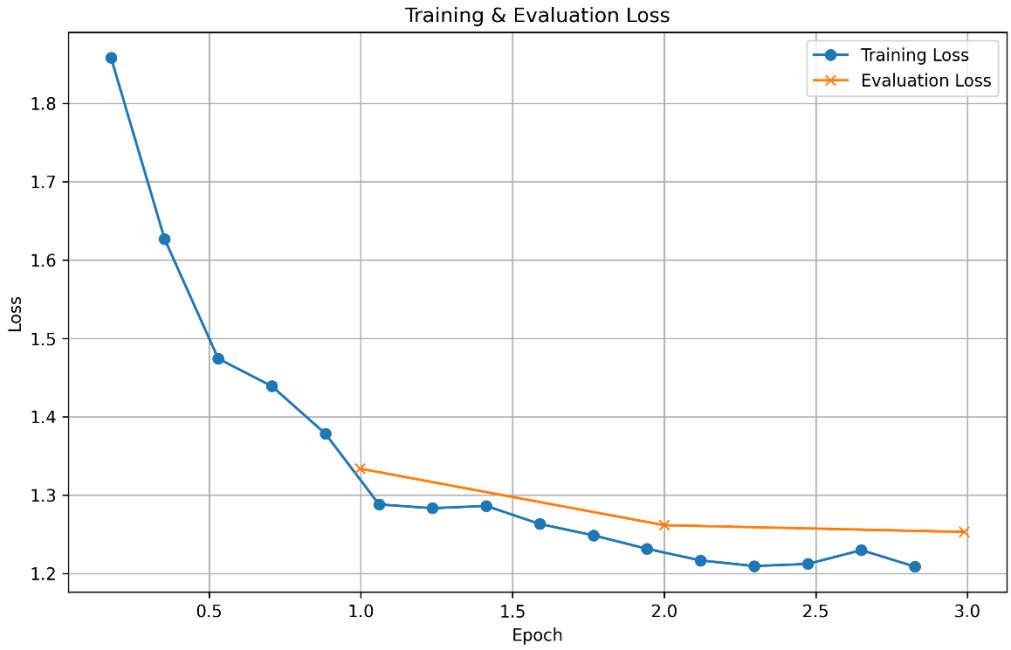
"question": "Bức ảnh mô tả những gì liên quan đến di sản văn hóa cách mạng?",
"prediction": "Bức ảnh mô tả một di sản văn hóa cách mạng, bao gồm các tài liệu lịch sử và hiện vật liên quan đến cuộc kháng chiến.",
"reference": "Bức ảnh cho thấy các bảng trưng bày tài liệu và hình ảnh liên quan đến Tổng Tấn Công và Nổi Dậy Xuân Mậu Thân 1968 tại Cần Thơ."

"question": "Cổng di tích có những đặc điểm kiến trúc nào nổi bật?",
"prediction": "Cổng di tích có kiến trúc truyền thống với các họa tiết trang trí tinh xảo, thể hiện nét văn hóa đặc trưng của Việt Nam.",
"reference": "Cổng di tích có mái ngói đỏ và các linh vật hình rồng ở trên cùng."

Hình 26: Kết quả dự đoán thực tế mô hình ViT, PhoBERT và ViT5 trên tập kiểm thử

3.2.2. Kết quả huấn luyện mô hình Qwen2-VL

Hình 27 minh họa diễn biến giá trị training loss và validation loss của mô hình Qwen2-vl qua các epoch



Hình 27: Biểu đồ loss mô hình Qwen2-VL

Training loss giảm ổn định qua các epoch, thể hiện quá trình tối ưu hiệu quả và sự thích nghi của mô hình với dữ liệu đa phương thức. **Validation loss** giảm ở giai đoạn đầu và đạt giá trị thấp tại giữa quá trình huấn luyện trước khi dao động nhẹ ở các epoch sau, cho thấy dấu hiệu overfitting ở mức thấp. Tuy nhiên, khoảng cách giữa hai đường loss vẫn nằm trong giới hạn chấp nhận được, chứng tỏ mô hình duy trì khả năng tổng quát hóa tốt.

Kết quả trên tập kiểm thử được trình bày trong Bảng 16:

Bảng 16: Kết quả đánh giá hiệu suất mô hình Qwen2-VL trên tập kiểm thử

BLUE	ROUGE1	ROUGE2	ROUGEL	BERTSCORE-F1	CIDER	EM
0.11	0.60	0.30	0.43	0.80	0.76	0.00

Phân tích kết quả:

- BLEU và ROUGE giữ mức tương tự mô hình Modular, cho thấy hai kiến trúc đạt mức độ tương đương trong việc nắm bắt từ khóa và cấu trúc câu.
- BERTScore-F1 và CIDEr duy trì mức cao, phản ánh khả năng hiểu thông tin ngữ nghĩa tốt dù cấu trúc câu trả lời khác biệt so với tham chiếu.
- EM bằng 0.00 tiếp tục phản ánh đặc thù linh hoạt về mặt biểu đạt trong VQA.

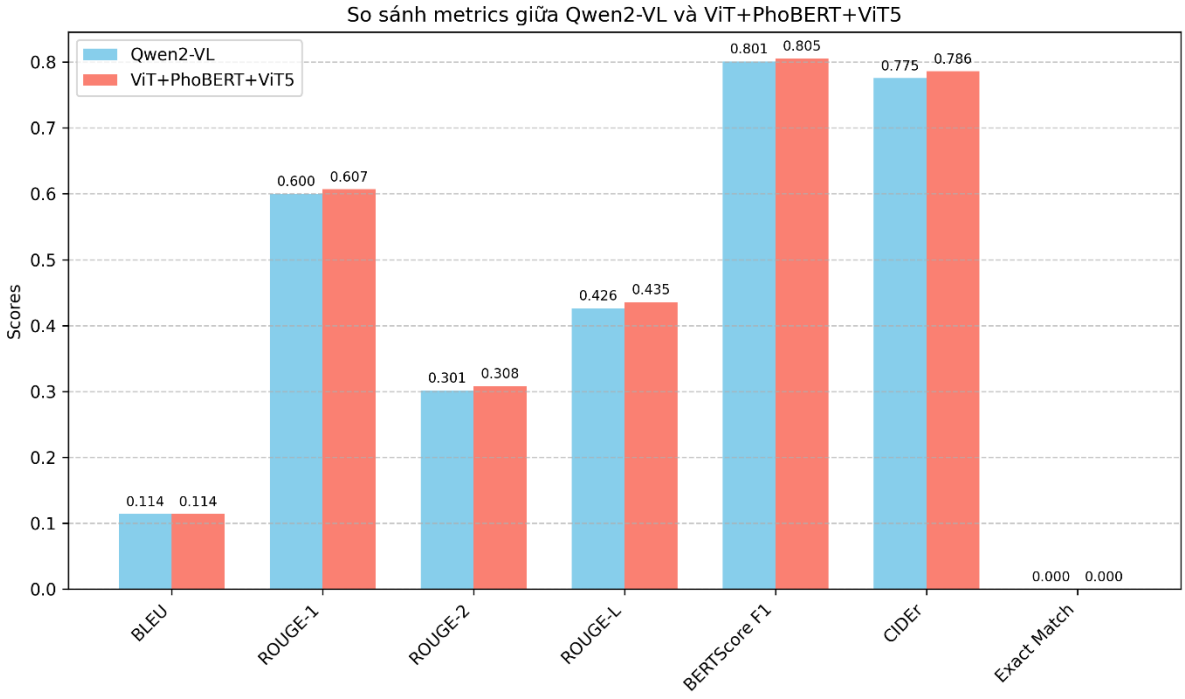
Hình 28 minh họa dự đoán thực tế, khẳng định mô hình có khả năng cung cấp câu trả lời phù hợp ngữ nghĩa với nội dung hình ảnh và câu hỏi.


```
"question": "Cổng vào di tích miếu Bà Chúa Xứ có những yếu tố kiến trúc nào nổi bật?",
"predicted": "Cổng vào được thiết kế với các họa tiết trang trí tinh xảo, thể hiện sự phong phú của văn hóa địa phương.",
"reference": "Cổng vào được trang trí với các họa tiết tinh xảo và màu sắc rực rỡ, bao gồm các tượng rồng và đèn lồng."
```


Hình 28: Kết quả dự đoán thực tế mô hình Qwen2-vl trên tập kiểm thử

3.2.3. So sánh kết quả 2 mô hình

Hình 29 và Bảng 17 thể hiện đối chiếu chi tiết giữa hai kiến trúc Modular và Unified:



Hình 29: Biểu đồ kết quả so sánh giữa kiến trúc Modular (ViT+PhoBERT+ViT5) và kiến trúc Unified (Qwen2-vl)

Bảng 17: So sánh kết quả giữa kiến trúc Modular (ViT+PhoBERT+ViT5) và kiến trúc Unified (Qwen2-vl)

Chỉ số đánh giá	ViT + PhoBERT+ViT5	Qwen2-vl	Chênh lệch
BLUE	0.11	0.11	0.00
ROUGE1	0.61	0.60	+0.01

ROUGE2	0.31	0.30	+0.01
ROUGEL	0.44	0.43	+0.01
BERTSCORE-F1	0.81	0.80	+0.01
CIDER	0.79	0.76	+0.03
EM	0.00	0.00	0.00

Nhận xét tổng hợp:

- Hai mô hình đạt hiệu suất tương đương về BLEU và ROUGE, chứng tỏ khả năng nhận diện từ khóa và cấu trúc câu tương đồng.
- Kiến trúc Modular đạt mức cao hơn về BERTScore-F1 và CIDEr, phản ánh khả năng nắm bắt ngữ nghĩa tốt hơn nhờ sử dụng các thành phần ngôn ngữ chuyên biệt cho tiếng Việt (PhoBERT, ViT5).
- Qwen2-VL duy trì hiệu suất cạnh tranh nhưng có ưu thế về tính tích hợp, giảm yêu cầu lắp ghép nhiều mô hình con.

3.3. Thảo luận về kết quả đạt được

Kết quả thực nghiệm cho thấy cả hai kiến trúc đều học hiệu quả các đặc trưng quan trọng từ dữ liệu hình ảnh và văn bản. Diễn biến loss ổn định chứng minh quá trình tối ưu được thực hiện đúng hướng và mô hình không gặp bất ổn trong quá trình hội tụ.

Mặc dù các chỉ số dựa trên n-gram như BLEU và ROUGE đạt giá trị thấp, đây không phải là hạn chế quan trọng đối với VQA, nơi câu trả lời đúng có thể được diễn đạt theo nhiều hình thức khác nhau. Do đó, các chỉ số ngữ nghĩa như BERTScore và CIDEr phản ánh chính xác hơn chất lượng của hệ thống.

So sánh hai kiến trúc cho thấy mô hình Modular có lợi thế về khả năng hiểu ngữ nghĩa, trong khi mô hình Unified mang lại một giải pháp tích hợp gọn nhẹ và hiệu quả, với hiệu suất chỉ giảm nhẹ. Điều này cho thấy Qwen2-VL là hướng tiếp cận khả thi cho các hệ thống VQA yêu cầu mô hình tổng quát và dễ triển khai, trong khi kiến trúc Modular phù hợp với bối cảnh tối ưu hóa cho tiếng Việt.

III. PHẦN KẾT LUẬN

1.1. Kết quả đạt được

Hệ thống trợ lý ảo ứng dụng Visual Question Answering (VQA) cho di sản văn hóa cách mạng Tây Nam Bộ đã được triển khai và đánh giá thành công trên hai kiến trúc: Modular (ViT + PhoBERT + ViT5) và Unified (Qwen2-VL). Kết quả thực nghiệm cho thấy nhiều đóng góp quan trọng.

Về hiệu quả mô hình, cả hai kiến trúc đều thể hiện khả năng học và biểu diễn tốt các đặc trưng từ dữ liệu hình ảnh và văn bản về di sản văn hóa. Quá trình huấn luyện diễn ra ổn định với training loss giảm đều từ 3.18 xuống 1.13 qua 13 epochs, trong khi validation loss duy trì ở mức hợp lý, chứng minh khả năng tổng quát hóa tốt. Các chỉ số ngữ nghĩa đạt giá trị cao, với BERTScore-F1 lần lượt là 0.81 và 0.80 cho Modular và Unified, CIDEr tương ứng 0.79 và 0.76, phản ánh mức độ tương đồng ngữ nghĩa tốt giữa câu trả lời sinh ra và câu tham chiếu, cho phép hệ thống hiểu chính xác nội dung câu hỏi ngay cả khi câu trả lời không trùng khớp hoàn toàn về mặt từ vựng.

Về so sánh kiến trúc, mô hình Modular thể hiện ưu thế về khả năng nắm bắt ngữ nghĩa tiếng Việt nhờ sử dụng các mô hình ngôn ngữ được tinh chỉnh chuyên biệt như PhoBERT và ViT5, phù hợp với bối cảnh tối ưu hóa chất lượng câu trả lời. Trong khi đó, kiến trúc Unified (Qwen2-VL) mang lại giải pháp gọn nhẹ và tích hợp cao hơn nhờ sử dụng chung một không gian biểu diễn cho cả hình ảnh và văn bản, đồng thời duy trì hiệu suất cạnh tranh với mức chênh lệch chỉ 0.01–0.03 điểm. Điều này khiến Unified trở thành lựa chọn phù hợp cho các hệ thống yêu cầu triển khai nhanh và dễ bảo trì.

Về ứng dụng thực tiễn, hệ thống đã chứng minh tính khả thi trong việc hỗ trợ người dùng tìm hiểu di sản văn hóa cách mạng Tây Nam Bộ theo cách trực quan và tương tác, góp phần nâng cao khả năng tiếp cận thông tin lịch sử – văn hóa trong không gian số. Bên cạnh đó, kết quả nghiên cứu cũng cung cấp cơ sở khoa học cho việc lựa chọn kiến trúc VQA phù hợp với từng mục tiêu triển khai: Modular cho yêu cầu chất lượng tối ưu, và Unified cho các kịch bản ưu tiên tính khả thi và khả năng mở rộng.

1.2. Hạn chế

Mặc dù đạt được nhiều kết quả tích cực, hệ thống trợ lý ảo vẫn tồn tại những hạn chế cần khắc phục trong các nghiên cứu tiếp theo.

Về chất lượng câu trả lời, chỉ số Exact Match (EM) đạt 0.00 trên cả hai mô hình, phản ánh hệ thống chưa tạo được câu trả lời hoàn toàn trùng khớp với tham chiếu. Mặc dù đây là đặc điểm phổ biến trong VQA do tính đa dạng trong cách diễn đạt, nhưng vẫn ảnh hưởng đến độ chính xác tuyệt đối trong một số trường hợp yêu cầu câu trả lời chuẩn. Các chỉ số dựa trên n-gram (BLEU = 0.11, ROUGE-2 = 0.30–0.31) còn thấp, cho thấy mô hình chưa tối ưu về cấu trúc câu, trật tự từ và khả năng tái tạo chuỗi từ liên tiếp – yếu tố quan trọng đối với câu trả lời dài hoặc mô tả chi tiết.

Về khả năng xử lý câu hỏi, hệ thống xử lý tốt các câu hỏi tổng quan (ví dụ: "Di tích này thuộc loại nào?") nhưng còn hạn chế với câu hỏi mang tính chi tiết cao như tên địa danh cụ thể, mốc thời gian chính xác, dữ kiện lịch sử chi tiết hoặc số liệu thống kê – những thông tin thường đòi hỏi truy xuất tri thức bên ngoài. Quy mô và độ đa dạng của tập dữ liệu hiện tại còn hạn chế, ảnh hưởng đến khả năng tổng quát hóa khi mô hình gặp hình ảnh hoặc câu hỏi mới nằm ngoài phân phối dữ liệu.

Về hiệu năng và trải nghiệm người dùng, validation loss xuất hiện dao động nhẹ trong giai đoạn sau của quá trình huấn luyện, cho thấy nguy cơ overfitting tiềm ẩn nếu mô hình được mở rộng mà không có các biện pháp điều chỉnh phù hợp (regularization, data augmentation). Thời gian phản hồi trung bình khoảng 8 giây cho mỗi câu hỏi, chưa đáp ứng kỳ vọng đối với ứng dụng trợ lý ảo tương tác thời gian thực, có thể gây trải nghiệm người dùng kém mượt mà. Hệ thống hiện chỉ hỗ trợ đối thoại đơn lượt (single-turn), chưa duy trì ngữ cảnh và lịch sử trao đổi giữa các câu hỏi liên tiếp, làm hạn chế khả năng xử lý hội thoại phức tạp và giảm tính tự nhiên trong tương tác.

1.3. Hướng phát triển

Để hoàn thiện hệ thống trợ lý ảo và mở rộng phạm vi ứng dụng, nghiên cứu có thể được phát triển theo các định hướng ưu tiên sau:

Tích hợp Retrieval-Augmented Generation (RAG): Kết hợp mô hình VQA với cơ chế truy xuất tri thức (knowledge retrieval) từ cơ sở dữ liệu cấu trúc về di sản văn hóa Tây Nam Bộ, bao gồm thông tin về di tích, sự kiện lịch sử, nhân vật, niên đại và số liệu thống kê. RAG giúp hệ thống sinh câu trả lời chi tiết, chính xác và giàu thông tin hơn bằng cách tham khảo nguồn tri thức bên ngoài, đặc biệt hiệu quả với các câu hỏi factual đòi hỏi dữ kiện cụ thể – điểm yếu chính của mô hình VQA truyền thống. Có thể áp dụng vector database (FAISS, Pinecone) kết hợp embedding model để truy xuất thông tin liên quan trước khi sinh câu trả lời, từ đó cải thiện cả độ chính xác và khả năng giải thích.

Phát triển hệ thống trợ lý ảo đa lượt hội thoại: Xây dựng cơ chế duy trì ngữ cảnh (context management) qua nhiều lượt tương tác, cho phép người dùng đặt câu hỏi follow-up mà không cần lặp lại thông tin. Tích hợp module quản lý hội thoại (dialogue management) để hỗ trợ các tính năng: làm rõ câu hỏi mơ hồ, gợi ý câu hỏi tiếp theo dựa trên ngữ cảnh, và tạo trải nghiệm tương tác tự nhiên, liền mạch. Phát triển giao diện người dùng thân thiện (chatbot interface, voice assistant) phù hợp với nhiều nhóm đối tượng: du khách, học sinh, giáo viên, và nhà nghiên cứu.

Ứng dụng trong môi trường giáo dục và du lịch văn hóa: Trong giáo dục, phát triển các tính năng hỗ trợ giảng dạy lịch sử như tạo bài kiểm tra tự động (quiz generation), gợi ý câu hỏi theo cấp độ kiến thức, kết nối thông tin di sản với bản đồ địa phương tương tác. Trong du lịch, tích hợp hệ thống vào ứng dụng di động hỗ trợ du

khách tại các điểm di tích, cung cấp thông tin theo ngữ cảnh (location-based information) và hỗ trợ đa ngôn ngữ (tiếng Việt, tiếng Anh). Phát triển tính năng AR (Augmented Reality) kết hợp VQA để tạo trải nghiệm tương tác phong phú hơn tại các di tích thực tế.

Mở rộng và tối ưu hóa dữ liệu: Tiếp tục thu thập, làm giàu và chuẩn hóa dữ liệu về di tích, nhân vật, sự kiện lịch sử; áp dụng kỹ thuật data augmentation (paraphrasing, image transformation) để tăng tính đa dạng và quy mô tập huấn luyện. Xây dựng quy trình chú thích (annotation) chất lượng cao với sự tham gia của chuyên gia lịch sử, đảm bảo tính chính xác và đầy đủ của thông tin tham chiếu. Cân nhắc mở rộng sang các khu vực di sản khác hoặc xây dựng tập dữ liệu chuẩn (benchmark) cho VQA tiếng Việt trong lĩnh vực văn hóa - lịch sử.

Tối ưu hóa hiệu năng và triển khai: Nghiên cứu áp dụng kỹ thuật tối ưu mô hình: model quantization, pruning, knowledge distillation để giảm kích thước mô hình và thời gian inference, hướng tới thời gian phản hồi < 3 giây. Triển khai hệ thống trên nền tảng cloud với khả năng scale động, hoặc tối ưu cho edge computing để hỗ trợ triển khai offline tại các điểm di tích.

Đánh giá và cải thiện liên tục: Thiết lập quy trình thu thập phản hồi từ người dùng thực tế (user feedback loop) để liên tục cải thiện chất lượng câu trả lời và trải nghiệm tương tác. Thực hiện đánh giá định tính từ chuyên gia lịch sử và giáo viên để đảm bảo tính chính xác về mặt học thuật và giá trị giáo dục.

TÀI LIỆU THAM KHẢO

[1] P. Bongini, F. Becattini, A. D. Bagdanov and A. Del Bimbo, "Visual Question Answering for Cultural Heritage," in *IOP Conference Series: Materials Science and Engineering*, 2020.

[2] F. Becattini, P. Bongini, L. Bulla, A. Del Bimbo, L. Marinucci, M. Mongiovì and V. Presutti, "VISCOUNTH: A Large-scale Multilingual Visual Question Answering Dataset for Cultural Heritage," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 6, pp. 1-20, 2023.

[3] W. Kim, B. Son and I. Kim , "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[4] J. Li, D. Li, C. Xiong and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proceedings of the 39th International Conference on Machine Learning*, 2022.

[5] J. Lu, D. Batra, D. Parikh and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

[6] H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," arXiv, 2019.

[7] K. V. Tran, K. V. Nguyen and N. L. T. Nguyen, "BARTPhoBEiT: Pre-trained Sequence-to-Sequence and Image Transformers Models for Vietnamese Visual Question Answering," in *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 2023.

[8] P. B. Nguyen, N. H. Tran and T. T. Quan, "ViCAN: Co-Attention Network for Vietnamese Visual Question Answering," in *Kỷ yếu Hội nghị KHCN Quốc gia lần thứ XIV về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR)*, Ho Chi Minh.

[9] S. N. Nguyen, S. V. Nguyen and T. Le, "Advancing Vietnamese Visual Question Answering with Transformer and Convolutional Integration," *Computers and Electrical Engineering*, vol. 119, p. 109474, 2024.

[10] H. Salehinejad, S. Sankar, J. Barfett, E. Colak and S. Valaee, "Recent Advances in Recurrent Neural Networks," 2018.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, "Attention Is All You Need," 2017.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," *arXiv preprint*, 2020.

[13] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[14] D. Q. Nguyen and A. T. Nguyen, "PhoBERT: Pre-trained language models for Vietnamese," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[15] L. Phan, H. Tran , H. Nguyen and T. H. Trinh, "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*, 2022.

[16] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou and J. Lin, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," 2024.

[17] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang and W. Chen, "LoRA: Low-Rank Adaptation of," 2021.

[18] K. Papineni, S. Roukos, T. Ward and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," 2002.

[19] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," 2004.

[20] R. Vedantam, L. C. Zitnick and D. Parikh, "CIDEr: Consensus-based Image Description Evaluation," 2015.

[21] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," 2019.

[22] P. Rajpurkar, J. Zhang, K. Lopyrev and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," 2016.