

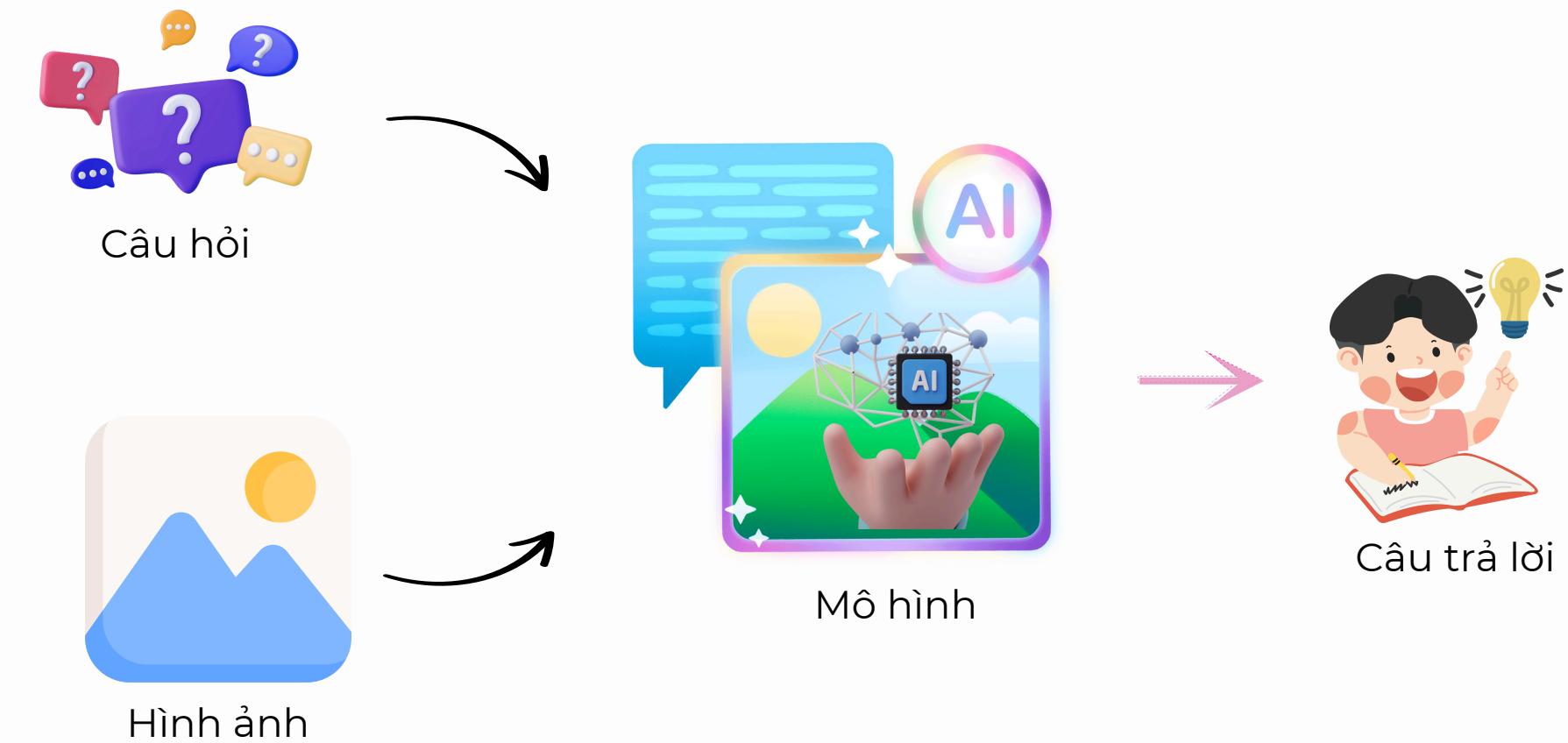
# Xây dựng hệ thống Visual Question Answering cho tiếng Việt

By: Nhóm 3



# Introduction

- Visual Question Answering (VQA) là bài toán yêu cầu khả năng kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên.



- Mục tiêu của VQA là dạy máy hiểu nội dung của hình ảnh và hiểu câu hỏi bằng văn bản sau đó trả lời các câu hỏi về hình ảnh đó bằng ngôn ngữ tự nhiên.

# Problems

Trong khi các bộ dữ liệu và mô hình VQA tiếng Anh đã được nghiên cứu rộng rãi, lĩnh vực này còn khá mới mẻ đối với tiếng Việt - một ngôn ngữ ít tài nguyên hơn.

## Low-resource Language

Lĩnh vực Visual Question Answering (vQA) cho tiếng Việt chưa được nghiên cứu nhiều và còn thiếu các bộ dữ liệu, mô hình chuyên biệt so với tiếng Anh.

## Computational Cost

Các mô hình VQA thường phức tạp và yêu cầu tài nguyên tính toán lớn, gây khó khăn cho việc triển khai trên các hệ thống hạn chế.

## Multimodal Fusion

Cần kết hợp hiệu quả thông tin từ hai nguồn khác nhau (hình ảnh và văn bản) để hiểu và trả lời câu hỏi.



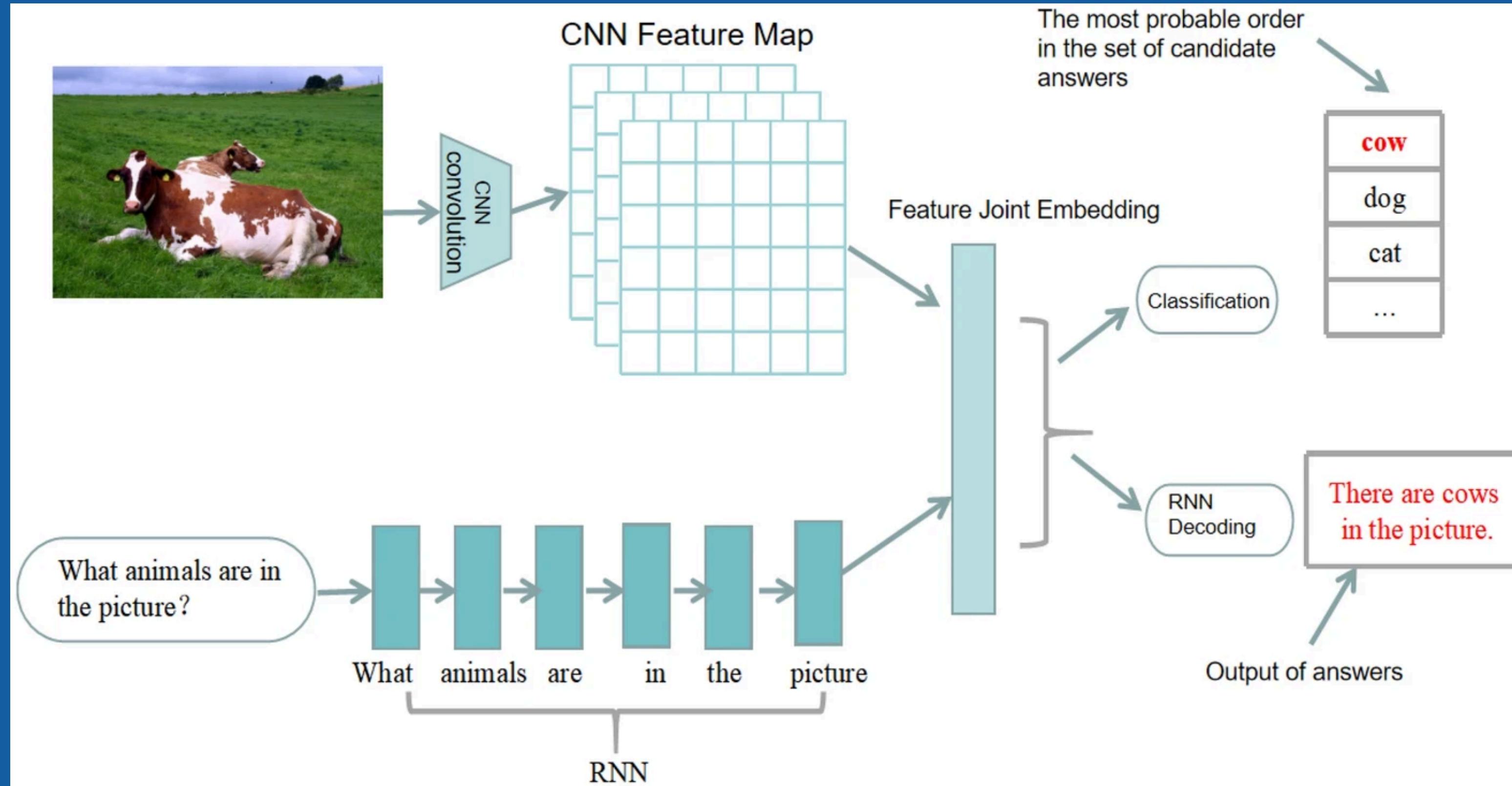
# Approaches

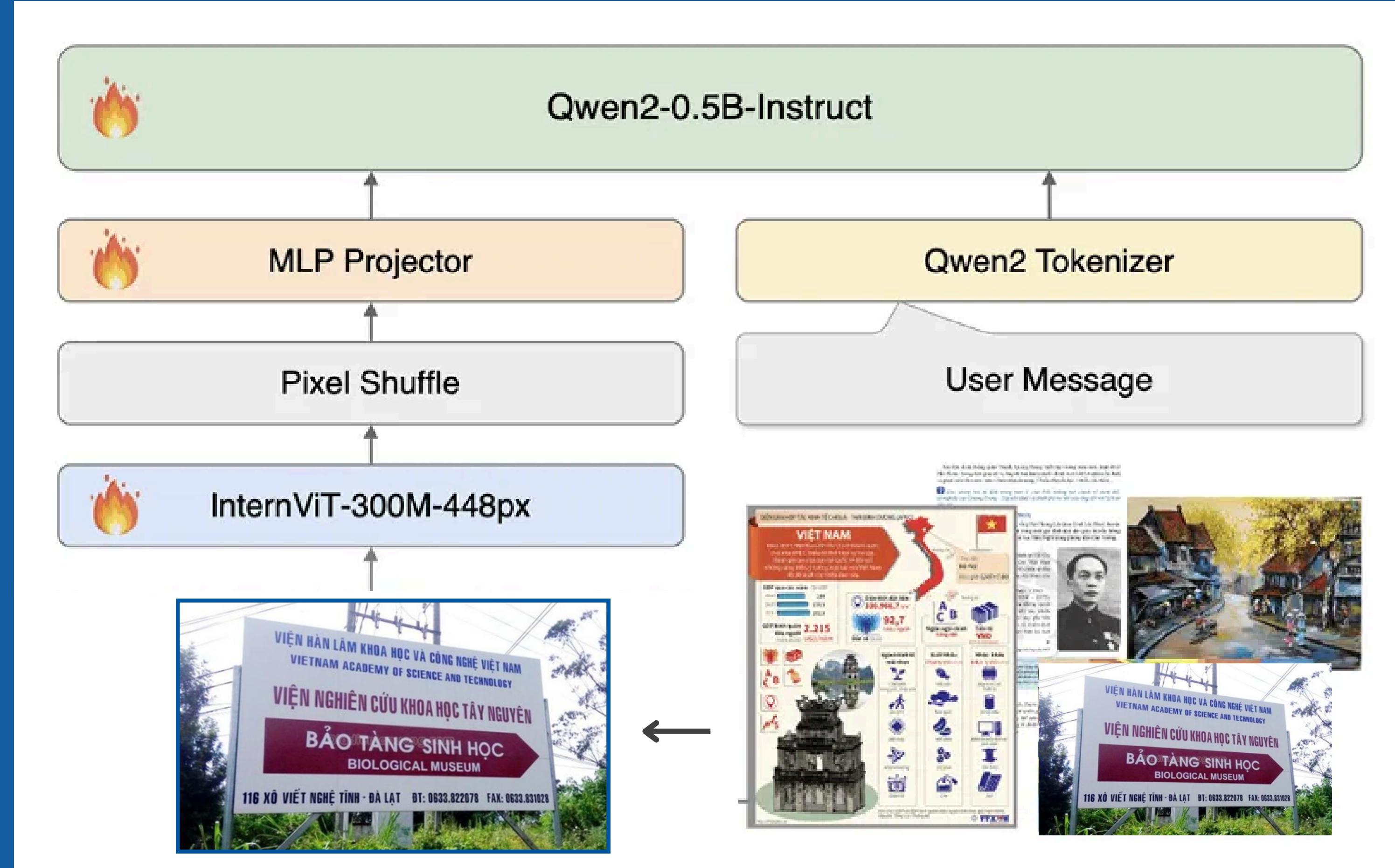
CNN và các  
biến thể +  
RNN (LSTM)

Các mô hình  
dựa trên cơ  
chế Attention  
( ViT +  
Transformer)

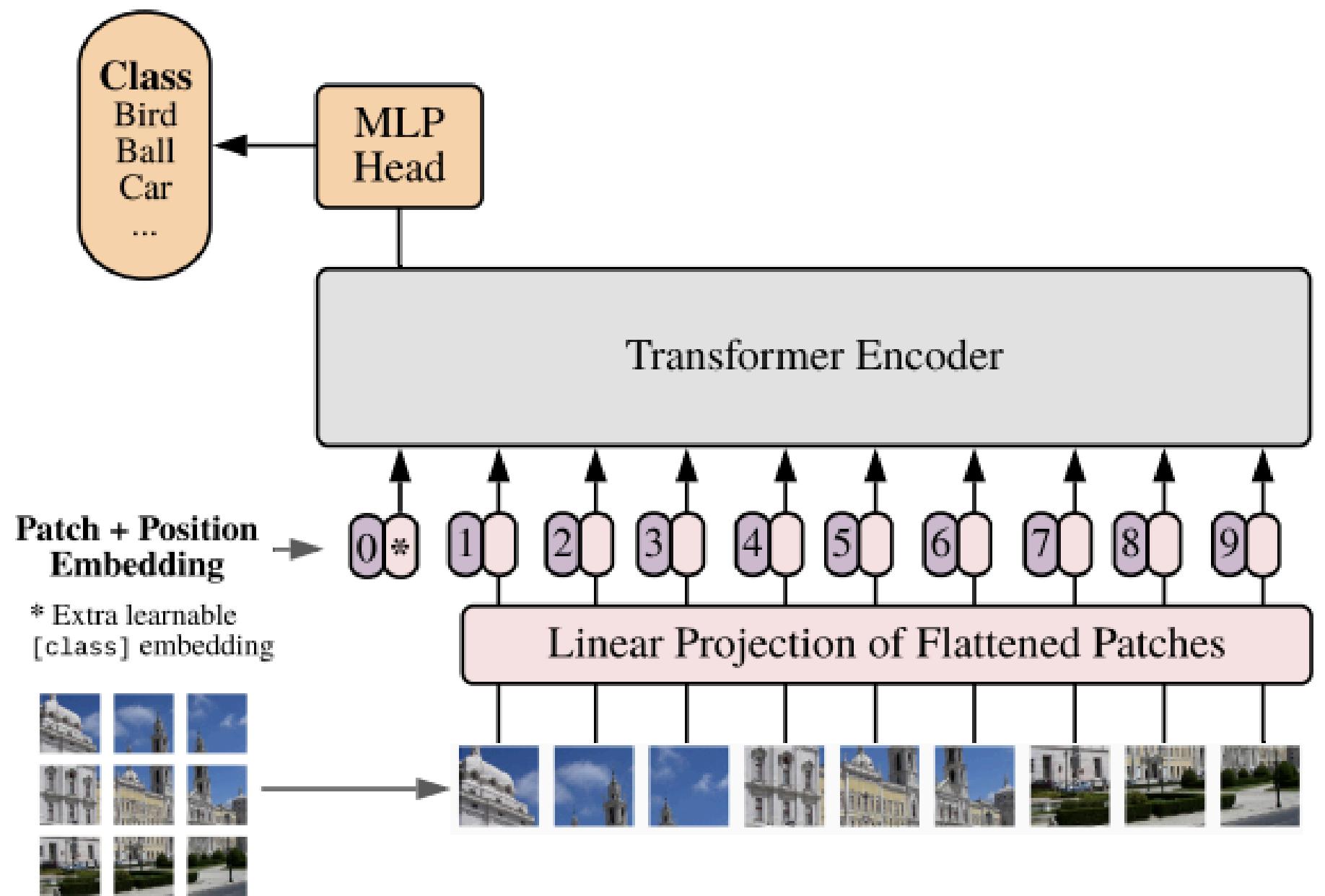
Mô hình tổng  
hợp (Unified  
Vision-  
Language  
Models)







# Vision Transformer (ViT)



vision\_model

vision\_model.embeddings

vision\_model.embeddings.class\_embedding

[1, 1, 1 024]

BF16

vision\_model.embeddings.patch\_embedding.bias

[1 024]

BF16

vision\_model.embeddings.patch\_embedding.weight

[1 024, 3, 14, 14]

BF16

vision\_model.embeddings.position\_embedding

[1, 1 025, 1 024]

BF16

448

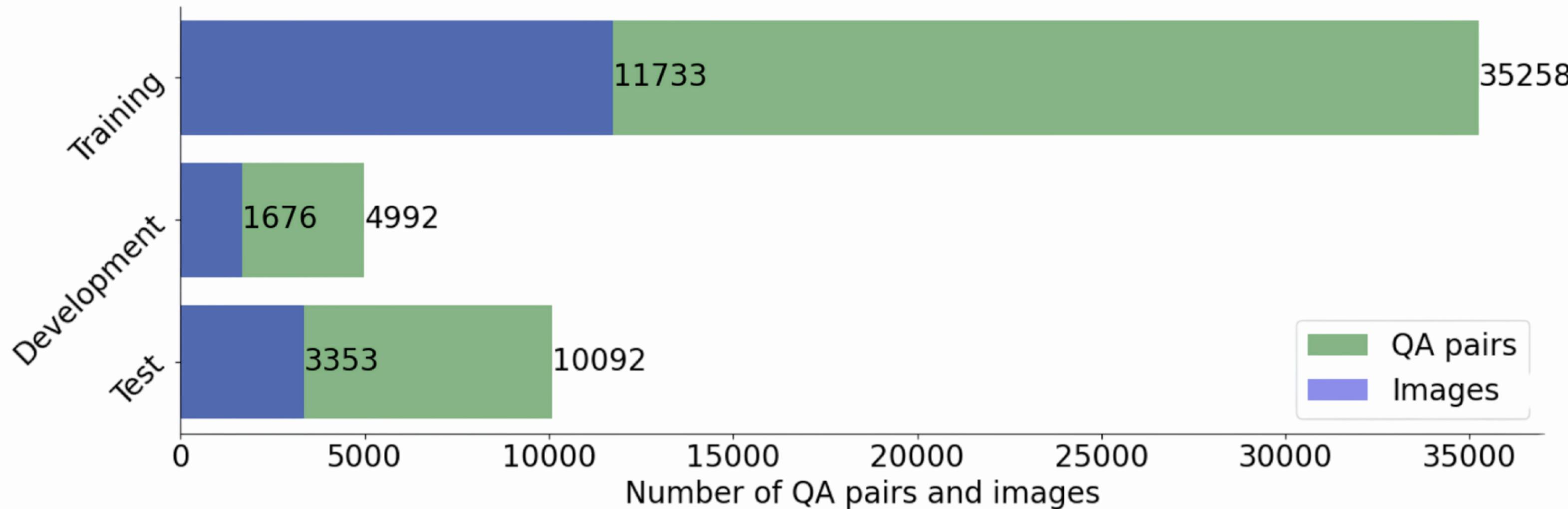
Báo cáo về sự kiện diễn ra tại Canada và Mỹ. Báo cáo về sự kiện diễn ra tại Canada và Mỹ.							
Form Phân tích Thuần Text							
Nhiều dữ liệu phân tích dữ liệu và các công cụ của sự kiện.							
Đến người liên quan: Bằng Khoa, Mỹ Văn.							
Ngày: Ngày 10/10/2023							
Một số tin tức chính phủ: Thủ tướng, Cảnh sát,							
Tóm tắt ảnh hưởng: Sự việc liên quan đến sự xuất hiện của ca sĩ Bằng Khoa tại Canada và việc hòa nhập Mỹ Văn. Đón nhận sự kiện này đã gây xôn xao dư luận trong cộng đồng người Việt tại Canada và Việt Nam, đặc biệt sau khi thông tin về sự cố với nhóm nhạc của ca sĩ Bằng Khoa được đưa ra. Phản hồi từ phía Mỹ Văn cho thấy có cảm thấy rõ truyền thông và sự phản ứng của người dân Canada và Việt Nam.							
Ghi chú sự kiện: (Thêm một nét về việc đón nhận sự kiện này. Ngày 10/10, bài phỏng vấn với Bằng Khoa, phản hồi của người dân tại Canada và Việt Nam.)							
Đánh giá: Đánh giá của Bằng Khoa: Ông có nhận xét tích cực về sự kiện, và tin tưởng nó sẽ mang lại nhiều niềm vui cho khán giả tại Mỹ.							
Đánh giá của Mỹ Văn: Ông đánh giá Bằng Khoa tại sân bay: Ông tin rằng sự thương kính với Bằng Khoa sẽ giúp đỡ, đặc biệt là bài viết trên báo của mình.							
Tóm tắt phản hồi truyền thông: Các phương tiện truyền thông tại Canada và Việt Nam đều có những bài viết, bình luận và sự kiện.							
Đánh giá của người dân:							
+ <b>Yêu cầu truyền thông:</b> Phóng viên đưa tin về sự kiện diễn ra tại Canada và Mỹ. Họ yêu cầu sự minh bạch và rõ ràng về thông tin phòng hộ của lực lượng an ninh.							
+ <b>Yêu cầu xã hội:</b> Sự quan tâm của người dân, sự đánh giá và sự kiện của người xem về mối quan hệ giữa hai							
Đánh giá của chính phủ:							
+ <b>Truyền thông:</b> Cần thận trọng hơn trong việc chia sẻ, tránh tiêu lầm và sụy diễn, đặc biệt trong vấn đề liên quan đến thông tin từ các cơ quan.							
Báo Mạng/Đài Phát thanh							
Ngày: Ngày 10/10/2023	Thời gian: 10:00 AM	Bằng Khoa: Bằng Khoa, Mỹ Văn	Sân bay: Sân bay quốc tế Toronto	Địa điểm: Canada, thành phố Toronto, Mỹ Văn	Nhịp sống: Sự việc thu hút sự chú ý của truyền thông và cư dân	Tin tức: Tin tức về sự kiện diễn ra tại sân bay	Người tham khao: Người dân Canada
16/10/2023	11:00 AM	Phú Quốc: Phú Quốc, Mỹ Văn	Phú Quốc: Phú Quốc, Mỹ Văn	Địa điểm: Phú Quốc, Mỹ Văn	Nhịp sống: Sự việc thu hút sự chú ý trong giới truyền thông	Tin tức: Tin tức về sự kiện diễn ra tại Phú Quốc	Người tham khao: Người dân Phú Quốc
17/10/2023	0:00 AM	François: François, Mỹ Văn	Quốc gia: Quốc gia	Địa điểm: Quốc gia	Nhịp sống: Sự kiện diễn ra tại Quốc gia	Tin tức: Tin tức về sự kiện diễn ra tại Quốc gia	Người tham khao: Chính phủ và nhân dân Quốc gia
17/10/2023	12:00 PM	Đưa tin của các báo: Các báo Việt Nam	Địa điểm: Các báo Việt Nam	Nhịp sống: Tin tức về sự kiện diễn ra tại các báo	Tin tức: Tin tức về sự kiện diễn ra tại các báo	Tin tức: Tin tức về sự kiện diễn ra tại các báo	Người tham khao: Các báo điện tử
Lưu ý: Thông tin nêu trên là ví dụ, cần bổ sung thêm chi tiết cụ thể và liên kết giữa các phần để áp dụng cho hoàn cảnh.							

448

# Dataset

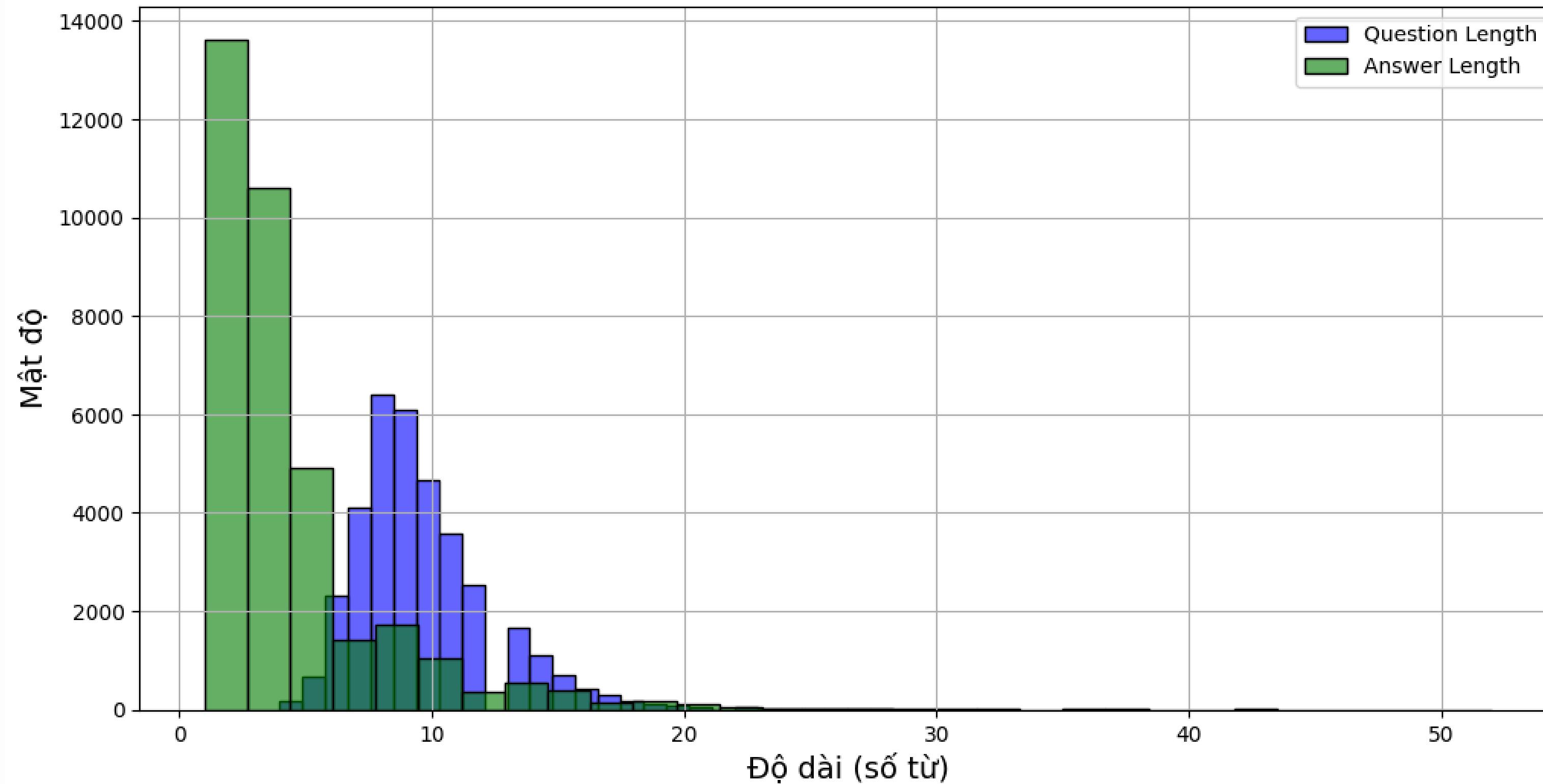
Dữ liệu	Có văn bản	Nguồn hình ảnh	Phương pháp chú thích	Ngôn ngữ	Thời gian xuất bản
ViVQA	Không	MS COCO	Bán tự động	Tiếng Việt	2021
EJVVQA	Không	Internet	Thủ công	Tiếng Anh, tiếng Nhật, tiếng Việt	2023
OpenViVQA	Có và không	Internet	Thủ công	Tiếng Việt	2023
ViTextVQA	Có	Internet, thủ công	Thủ công	Tiếng Việt	2024

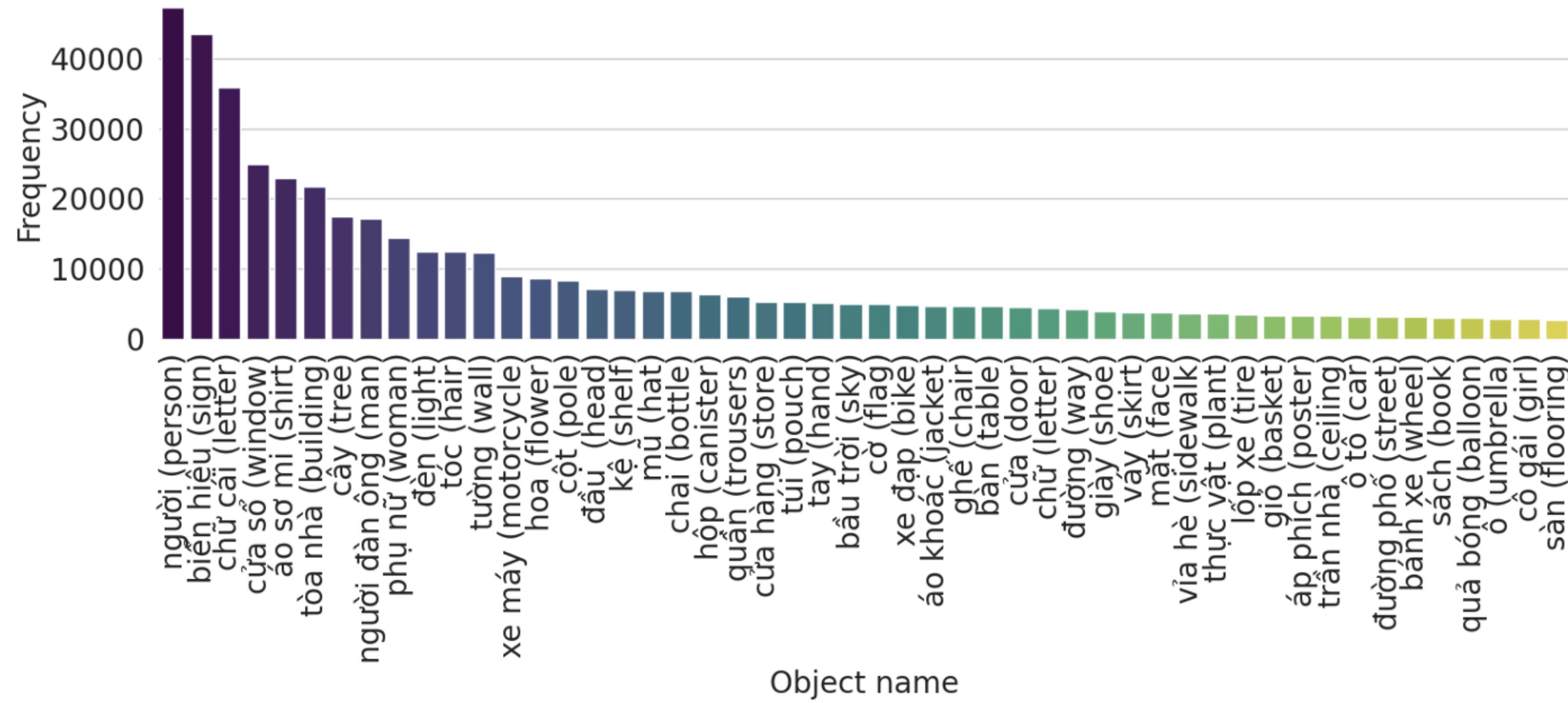
## ViTextVQA dataset



# ViTextVQA dataset

Phân phối độ dài của Question và Answer





Phân bố các đối tượng chính trong dataset.

# Thực nghiệm

- Dataset : ViTextVQA
- Sử dụng mô hình kết hợp CNN (và biến thể) + RNN (LSTM) cho trích xuất đặc trưng và sinh câu trả lời → kết quả không tốt.
- Training từ đầu với tập dữ liệu nhỏ → không hiệu quả.
- Fine-tune mô hình Vintern-1B: Tận dụng pretraining từ tập dữ liệu lớn, giảm tài nguyên huấn luyện.
- Kỹ thuật tối ưu tài nguyên:
  - Đóng băng: Vision Encoder và MLP Projector.
  - Fine-tune: LLM với LoRA (rank=16) (trainable params: 8,798,208)
  - Sử dụng bf16 thay vì float32.
  - Resize ảnh cố định 448x448 (thay vì chia ô dynamic high-resolution).

# Evaluation

- **Exact Match**

$$EM = \frac{\text{Số lượng câu trả lời dự đoán đúng hoàn toàn}}{\text{Tổng số câu trả lời}}$$

- **F1-Score**

$$\text{Precision} = \frac{\text{Số lượng token dự đoán đúng}}{\text{Tổng số token được dự đoán}}$$

$$\text{Recall} = \frac{\text{Số lượng token dự đoán đúng}}{\text{Tổng số token trong đáp án chuẩn}}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



# Kết quả

Mô hình	F1 Score (%)	EM (%)
BLIP-2	38,47%	15,58%
PreSTU	44,93%	22,64%
ViTextBLIP-2	53,95%	25,48%
<b>Ours</b>	<b>62,79%</b>	<b>40,73%</b>

Link submit: <https://www.kaggle.com/competitions/ViTextVQA-evaluation>

# Kết luận:

- Đạt kết quả khá tốt, nhưng hiện chỉ đang tập trung trên tập ViTextVQA thôi và đang phụ thuộc vào pretrained.
- Có thể huấn luyện với tài nguyên mở (kaggle, gg colab)

## Hướng phát triển :

- Có thể cải tiến dữ liệu để có văn phong tự nhiên hơn
- Giảm kích thước mô hình

thank you!

