

RL for Multimodal Reasoning in MLLMs

Phạm Thế Anh

| TABLE OF CONTENTS

- 1. Bối Cảnh**
- 2. Mục Tiêu**
- 3. Kiến Thức Cốt Lõi**
- 4. Pipeline Triển Khai**
- 5. Kết Quả Huấn Luyện**
- 6. Trước & Sau Fine-tune**
- 7. Kết Luận**
- 8. Hướng Phát Triển**



TRƯỜNG ĐẠI HỌC FPT

Bối Cảnh

REL301m

QUY NHON A.I CAMPUS

|1. BỐI CẢNH

- **Các mô hình ngôn ngữ đa phương thức (MLLMs) như LLaVA, Qwen-VL có thể xử lý ảnh + văn bản.**
- **Thách thức: Suy luận đa bước từ hình ảnh, đặc biệt với câu hỏi trừu tượng, logic.**
- **Mô hình sử dụng: Qwen-VL-2B-Instruct – open-source, nhẹ, hỗ trợ cả ảnh & text.**



TRƯỜNG ĐẠI HỌC FPT

Mục Tiêu

REL301m

QUY NHON A.I CAMPUS

| 2. MỤC TIÊU

- Áp dụng RLHF/DPO để huấn luyện mô hình biết “chọn” câu trả lời tốt hơn.
- Dữ liệu gồm: **image_path, prompt, chosen, rejected**.

```
"image_path": "/content/drive/MyDrive/REL_Project/Data/train/images/aguanambi-1085_png_jpg.rf.60d109572",
"prompt": "<|user|>\nPlease count the number of vehicles you see. <image>\n<|assistant|>\n",
"chosen": "There are 2 cars, 3 motorcycles, 0 buss, 0 bicycles in the image.",
"rejected": "There are 5 cars, 6 motorcycles, 3 buss, 9 bicycles in the image."
```

- Sử dụng DPO (Direct Preference Optimization) thay cho PPO:
 - Đơn giản, không cần reward model.
 - Giảm chi phí huấn luyện và ổn định hơn.



TRƯỜNG ĐẠI HỌC FPT

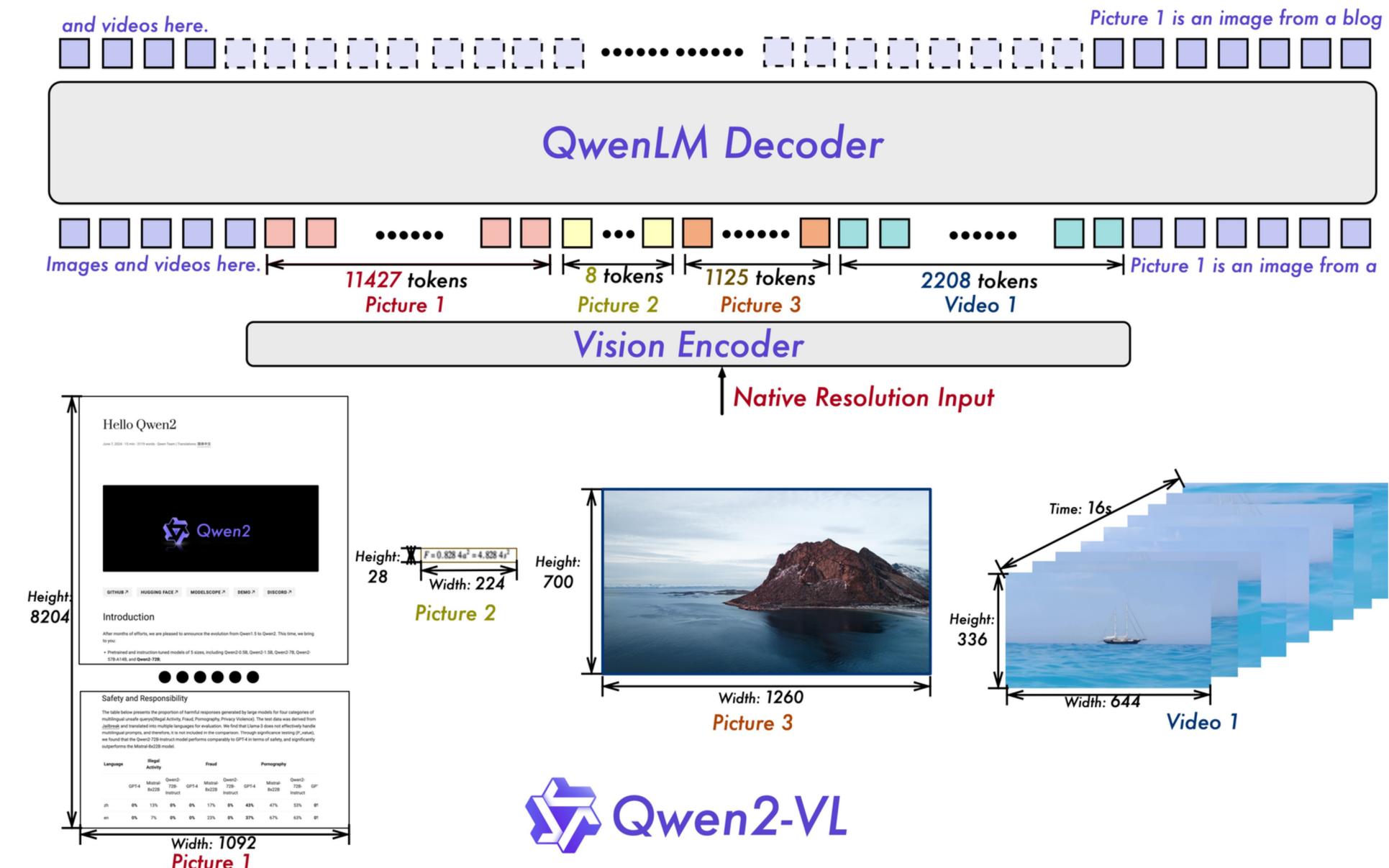
Kiến Thức Cốt Lõi

REL301m

QUY NHON A.I CAMPUS

| 3. KIẾN THỨC CỐT LÕI

- **Qwen-VL-2B-Instruct là một mô hình ngôn ngữ đa phương thức với:**
 - **2 tỷ tham số**
 - Kiến trúc **Transformer decoder-only**
 - Hỗ trợ xử lý đồng thời văn bản và hình ảnh (Visual + Language)



| 3. KIẾN THỨC CỐT LÕI

Thành phần chính:

- **Vision Encoder:** ViT (Vision Transformer) trích xuất đặc trưng từ ảnh.
- **Q-Former:** Giao tiếp giữa đặc trưng hình ảnh và token ngôn ngữ.
- **Tokenizer:** Hỗ trợ cả hình ảnh (<image> token) và văn bản tự nhiên.
- **Instruct-tuned:** Được huấn luyện với instruction dataset (như LLaVA, RefCOCO, MME,...).

Khả năng:

- Trả lời câu hỏi về ảnh
- Mô tả nội dung ảnh (captioning)
- Thực hiện suy luận đa bước trên nhiều modality

| 3. KIẾN THỨC CỐT LÕI

DPO là gì?

- Phương pháp học từ phản hồi đánh giá (ranking hoặc pairwise feedback)
- Không cần **reward model** như RLHF truyền thống
- Không cần sinh mẫu mới trong lúc train → hiệu quả cao hơn

Công thức Loss (giản lược):

$$L = -\log \sigma(\Delta)$$

Trong đó:

- $\Delta = \text{logit(preferred)} - \text{logit(rejected)}$
- σ là hàm sigmoid

| 3. KIẾN THỨC CỐT LÕI

Ưu điểm:

- Không cần sample rollout
- Training ổn định hơn PPO
- Phù hợp cho mô hình lớn và dữ liệu giới hạn

✓ Thư viện sử dụng:

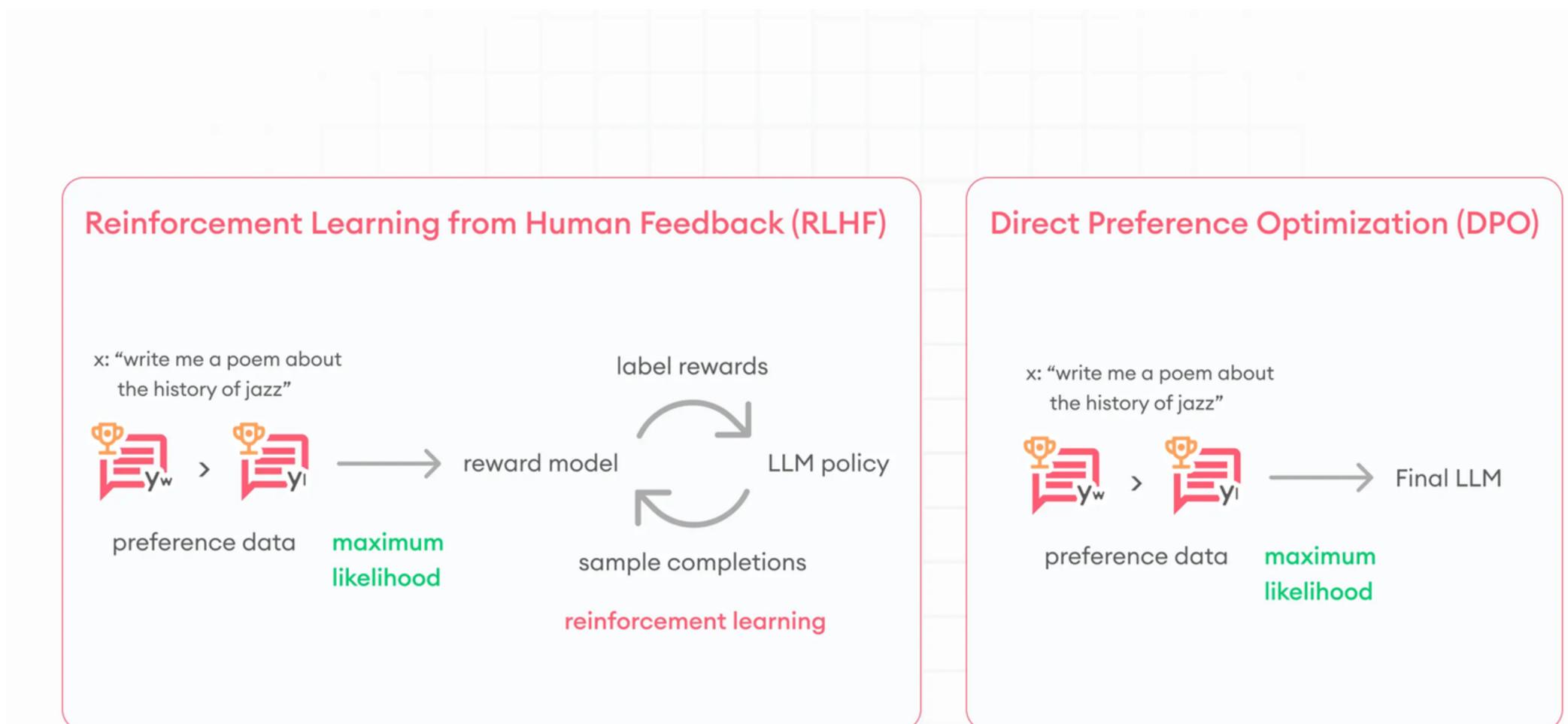
- trl từ HuggingFace: hỗ trợ DPOTrainer
- Tương thích tốt với unsloth để nạp mô hình nhanh hơn

| 3. KIẾN THỨC CỐT LÕI

-RLHF: Học từ phản hồi của con người (gồm cả ranking/trả lời).

-DPO:

- Không cần reward model.
- Trực tiếp tối ưu hoá từ cặp preferred vs rejected.
- Loss: $-\log P(\text{preferred}) + \log P(\text{rejected})$





TRƯỜNG ĐẠI HỌC FPT

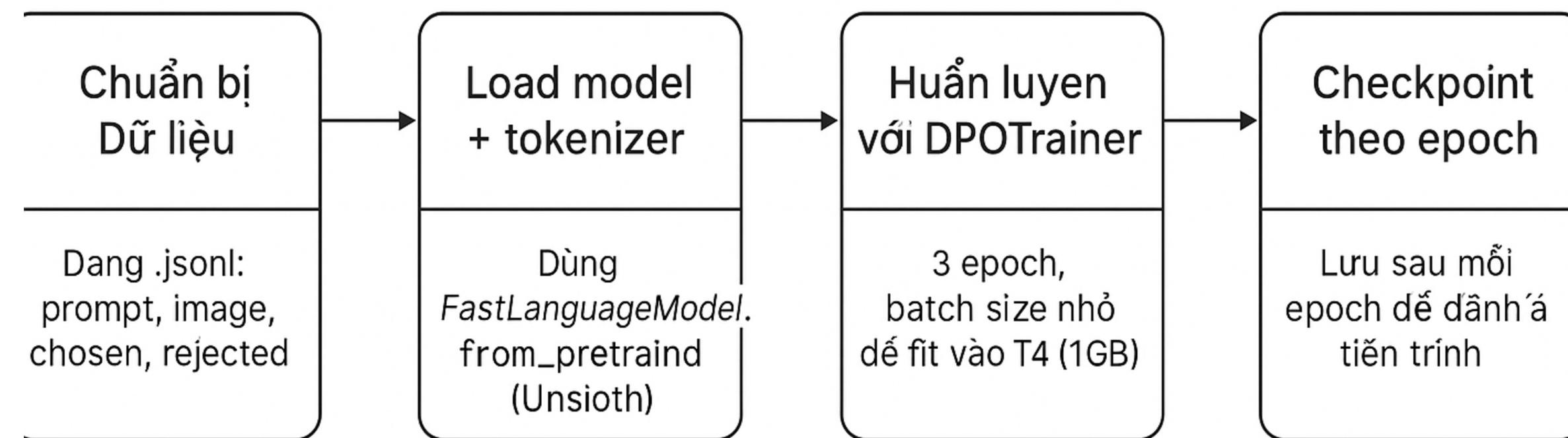
Pipeline Triển Khai

REL301m

QUY NHON A.I CAMPUS

|4. PIPELINE TRIỂN KHAI

Pipeline Triển Khai





TRƯỜNG ĐẠI HỌC FPT

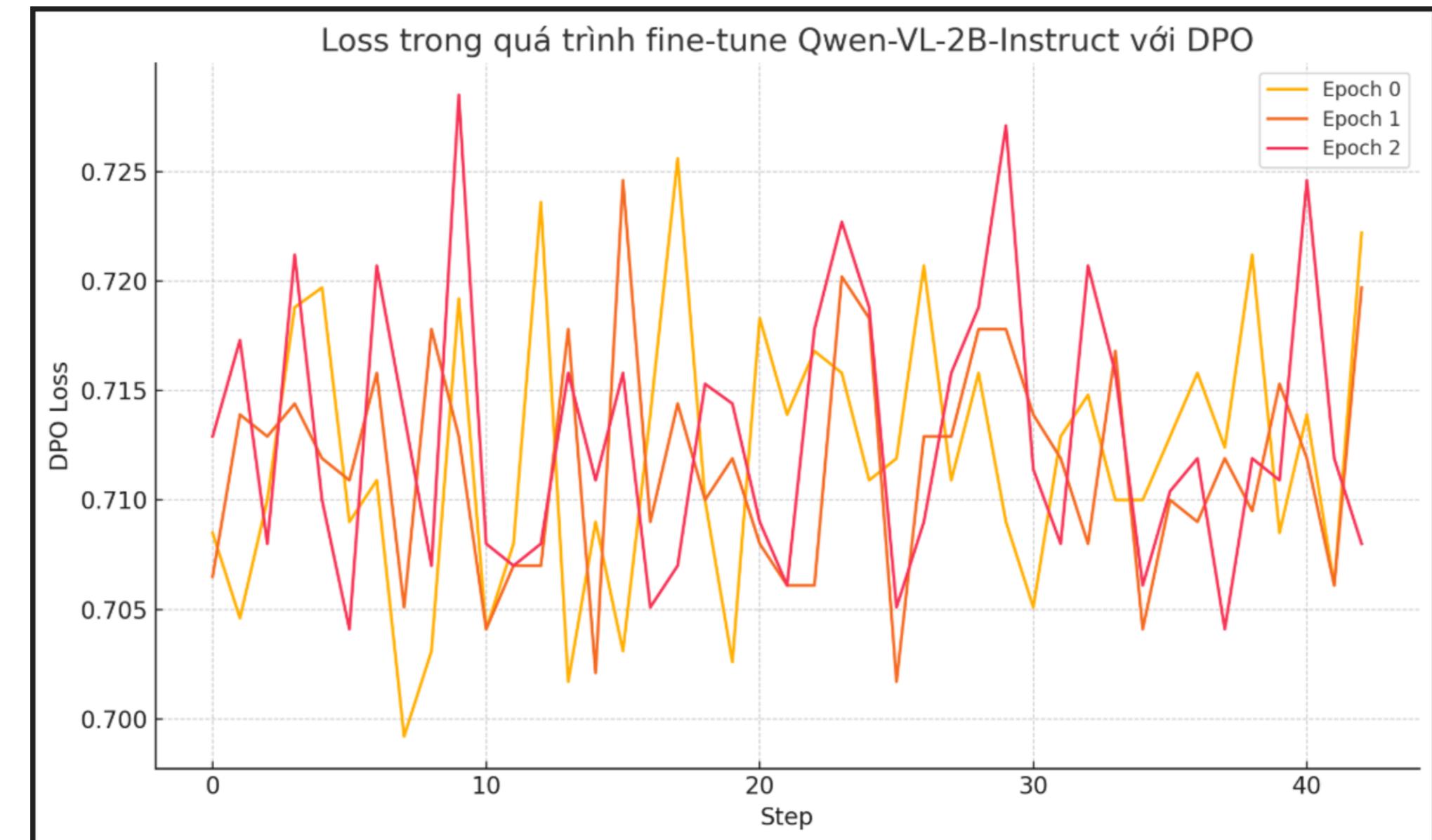
Kết Quả Huấn Luyện

REL301m

QUY NHON A.I CAMPUS

| 5. KẾT QUẢ HUẤN LUYỆN

- **Loss dao động ổn định quanh 0.70–0.72**
- **Tốc độ nhanh do model nhỏ (2B) + DPO không cần reward model**
- **Tổng số bước/epoch: ~420**





TRƯỜNG ĐẠI HỌC FPT

Trước & Sau Fine- tune

REL301m

QUY NHON A.I CAMPUS

| 6. TRƯỚC & SAU FINE-TUNE

Trước fine-tune (mô hình gốc) :

```
Using a slow image processor as `use_fast` is unset and a
You have video processor config saved in `preprocessor.js`
['There are 7 vehicles in the photo.']
```

Sau fine-tune :

```
Kết quả mô hình:
<|user|>
Can you tell me the total number of vehicles in this photo? <image>
<|assistant|>
There are 7 cars and 1 truck in the picture.
```



TRƯỜNG ĐẠI HỌC FPT

Kết Luận

REL301m

QUY NHON A.I CAMPUS

| 7. KẾT LUẬN

- DPO hoạt động tốt với MLLM mà không cần RL truyền thống.
- Qwen-VL-2B-Instruct là một nền tảng nhẹ, dễ fine-tune.
- Kết quả huấn luyện cho thấy loss ổn định và khả năng mô tả ảnh được cải thiện.
- Tuy nhiên:
 - Chưa đánh giá BLEU, ROUGE hoặc benchmark.
 - Vẫn cần nhiều dữ liệu hơn.



TRƯỜNG ĐẠI HỌC FPT

Hướng Phát Triển

REL301m

QUY NHON A.I CAMPUS

| 8. HƯỚNG PHÁT TRIỂN

- **Fine-tune thêm epoch đến khi loss hội tụ rõ rệt.**
- **Áp dụng LoRA để giảm yêu cầu bộ nhớ (RAM).**
- **Đánh giá định lượng:**
 - MME (Multimodal Eval)
 - LLaVA-Bench
 - Human Preference Study (nếu có).
- **Thêm số lượng data (khoảng 10.000 ảnh)**

DSP391m



DSP391m-G2

THANK YOU

FPT UNIVERSITY

QUY NHON A.I CAMPUS