# VIETNAMESE IMAGE CAPTIONING USING CLIP PREFIX AND GPT-2 LANGUAGE MODEL


Vu Nguyen Ha Anh[1]

[1]University of Information Technology – UIT, Ho Chi Minh City, Vietnam


## WHY?

- **Language Gap:** Lack of SOTA captioning solutions for Vietnamese (low-resource language).
- **Resource Efficiency:** Avoids high data and compute costs of training from scratch.

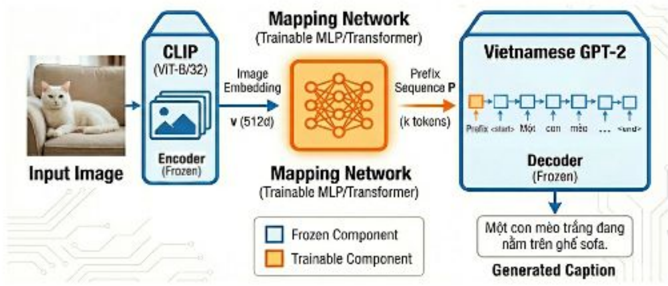

## WHAT?

**ViClipCap** introduced as a **lightweight framework** for Vietnamese Image Captioning based on **Prefix Tuning.** We have:

- **Architecture:** Bridges **frozen CLIP** and **VN GPT-2** via a lightweight **Mapping Network**.
- **Method:** Projects visual features into semantic prefixes for generation.
- **Impact:** Parameter-efficient transfer learning for low-resource languages.

## OVERVIEW

**Mechanism:** A **semantic translator** bridging frozen CLIP and GPT-2. Lightweight Mapping Network converts visual insights into language prompts without retraining backbones.




1. **Frozen CLIP:** Extracts fixed semantic visual embeddings.
2. **Trainable Mapping Network:** Projects visual features into semantic prefixes (The core innovation).
3. **Frozen VN GPT-2:** Autoregressively generates Vietnamese captions from the prefix context.

## METRICS

- **BLEU-4:** Measures precision of 4-gram overlaps.
- **ROUGE-L:** Focuses on recall via Longest Common Subsequence.
- **METEOR:** Aligns tokens using synonyms and stemming.
- **CIDEr:** Uses TF-IDF to weight consensus (caption-specific).
- **SPICE:** Evaluates semantic accuracy via Scene Graphs.

## DATASET

- **Selection:** Prioritized KTVIC (Life Domain) over UIT-ViIC (Sports) to capture diverse daily activities.
- **Focuses : Life Domain,** daily activities for Vietnamese context.
- **Scale: 4,327 images** annotated with **21,635 captions** (~5 captions/image).
- **Goal:** Addresses low-resource challenges in Vietnamese Vision-Language research.

## DESCRIPTION

### 1. Frozen Visual Encoder (CLIP)

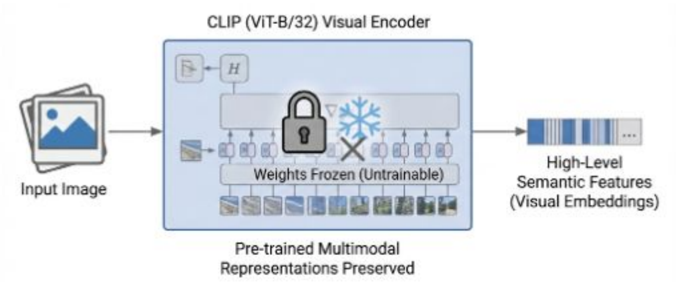



**Figure 1: Frozen Visual Encoder (CLIP).** ViT-B/32 backbone extracts features with locked weights, preserving pre-trained knowledge and reducing compute.

- **Backbone**: **CLIP (ViT-B/32)** extracts high-level semantic features.
- **Mechanism: Frozen weights** preserve pre-trained knowledge and significantly reduce compute.

### 2. Trainable Mapping Network

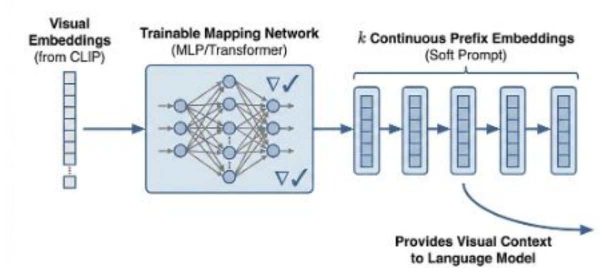



**Figure 2: Trainable Mapping Network.** Projects visual features into continuous prefix embeddings (soft prompts).

- **Bridge:** Links CLIP & GPT-2 via a lightweight Mapping Network.
- **Function:** Projects features into **Prefix Embeddings** ("soft prompts") to guide the **frozen LM**.

### 3. Frozen Text Decoder (Vietnamese GPT-2)

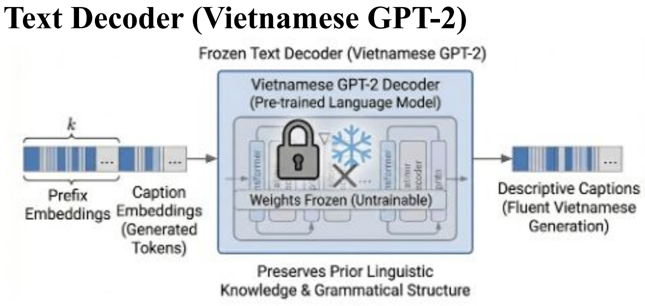



**Figure 3: Frozen Text Decoder (Vietnamese GPT-2).** Generates fluent captions from prefix inputs while keeping weights fixed.

- **Decoder:** Pre-trained **Vietnamese GPT-2** (Frozen).
- **Input:** Concatenated sequence:

  `[Prefix Embeddings, Caption Embeddings]`
- **Benefit:** Preserves linguistic knowledge for **fluent Vietnamese generation**.

### 4. Training Objective & Optimization

- **Objective**: Minimize **Cross-Entropy Loss**:

$$\mathcal{L} = -\sum_{t=1} \log P_\theta(y_t \mid y_{<t}, \mathbf{p})$$

- **Optimization: AdamW** with **Linear Warmup.**
- **Efficiency:** Updates *only* Mapping Network, preventing catastrophic forgetting.

### 5. Experimental Results

| Metric | CNN+LSTM | ViCLIPCap |
|---|---|---|
| BLEU-4 | 0.2572 | **0.3431** |
| ROUGE-L | 0.4895 | **0.5204** |
| CIDEr | 0.6282 | **0.8127** |
| METEOR | 0.2995 | **0.3194** |
| SPICE | 0.0782 | **0.0829** |

+ ViClipCap outperforms the CNN+LSTM across **all metrics** on the KTVIC dataset.
+ Demonstrates that high quality can be achieved with minimal trainable parameters via **Prefix Tuning**.

## CONCLUSION

- **Adaptation:** Optimized CLIP & GPT-2 for Vietnamese Life Domain (KTVIC).
- **Performance:** Fluent, efficient generation via Prefix Tuning without catastrophic forgetting.
- **Future:** Scaling to larger backbones and multilingual expansion.

## REFS


[1] Ron Mokady, Amir Hertz, Amit H. Bermano: ClipCap: CLIP Prefix for Image Captioning.

[2] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, Rita Cucchiara: From Show to Tell: A Survey on Deep Learning-Based Image Captioning.



Vu Nguyen Ha Anh
University of Information Technology

TEL : 0822654859
Email : vunha.20@grad.uit.edu.vn