Multimodal Video Summarisation

Alternative Vision and Language Backbones

Siddharth[1], Mohit[1], Yashasvi[1], Sahil Shukla[1], Navdeep[1] , Sanjay[1],
Indian Institute of Technology Mandi
(t25103,s25019,s25018,d25044,ptd2508)@students.iitmandi.ac.in

Problem Statement

- Although video summarization has advanced with deep learning and vision-language models (VLMs), most existing works are limited in scale, modality, and generalization. Traditional datasets and models primarily focus on video-to-video summarization and fail to fully leverage multimodal (video + text) understanding.
- This project addressed this by introducing a unified model capable of generating video and text summaries through temporal and task prompts. However, the original approach was implemented using LLaMA and CLIP, which restrict the model's language comprehension and visual-text alignment.
- Hence, there is a need to reproduce and enhance this framework using stronger large language models and more advanced vision encoders to improve semantic accuracy, multimodal alignment, and summary quality across different summarization tasks (V2V, V2T, and V2VT).

Methodology

Dataset: Activity Net-Cap: 20k multimodal video-text pairs

⇒ **Frame Extraction:**
Input video is sampled at 1 FPS to extract representative frames

$$F = \{f_1, f_2, \ldots, f_n\}$$

⇒ **Feature Extraction using CLIP:** Each frame f(i) is passed through a CLIP Vision Encoder to generate visual embeddings

$$v_i = E_v(f_i).$$

CLIP (Contrastive Language–Image Pretraining) jointly trains image and text encoders using a contrastive learning objective, enabling strong alignment between visual and semantic representations. The visual embeddings obtained from CLIP effectively capture high-level image features that are useful for downstream tasks.

⇒ **Temporal Representation:**
Frames are assigned temporal tokens such as [t1],[t2],[t3].....[tn] to maintain sequence information.
Combined representation:

$$S = \{t_1, v_1, t_2, v_2 \ldots, t_n, v_n\}$$

⇒ **Summarization via LLM Models:** The text decoder receives interleaved tokens and generates modality-controllable summaries:

$$A_x = \{LLM(S, I_x)\}$$

⇒ **Evaluation and Comparison:** The generated summaries are evaluated using quantitative metrics (BLEU, ROUGE, CIDEr) and qualitative analysis (factual consistency and visual alignment).

⇒ **Training Objective (Negative Log-Likelihood):**
The loss function for our video summarization model can be written as:

$$L = -\sum_{i=1}^{N} \log p(A_{x_i} \mid S, A_{x_{<i}}) \qquad (1)$$

where:

- L is the total loss.
- N is the total number of selected video segments or frames.
- A_{x_i} is the action (or frame/segment) predicted at step i.
- S is the input video sequence or context.
- $A_{x_{<i}}$ denotes all previous predicted actions (frames/segments) before step i.

Objectives

- To develop a cross-modal video summarization model combining visual and textual understanding.
- Provide results on new evaluation metrics (FCLIP Cross-FCLIP) that capture semantic similarity, not just exact matches

Workflow

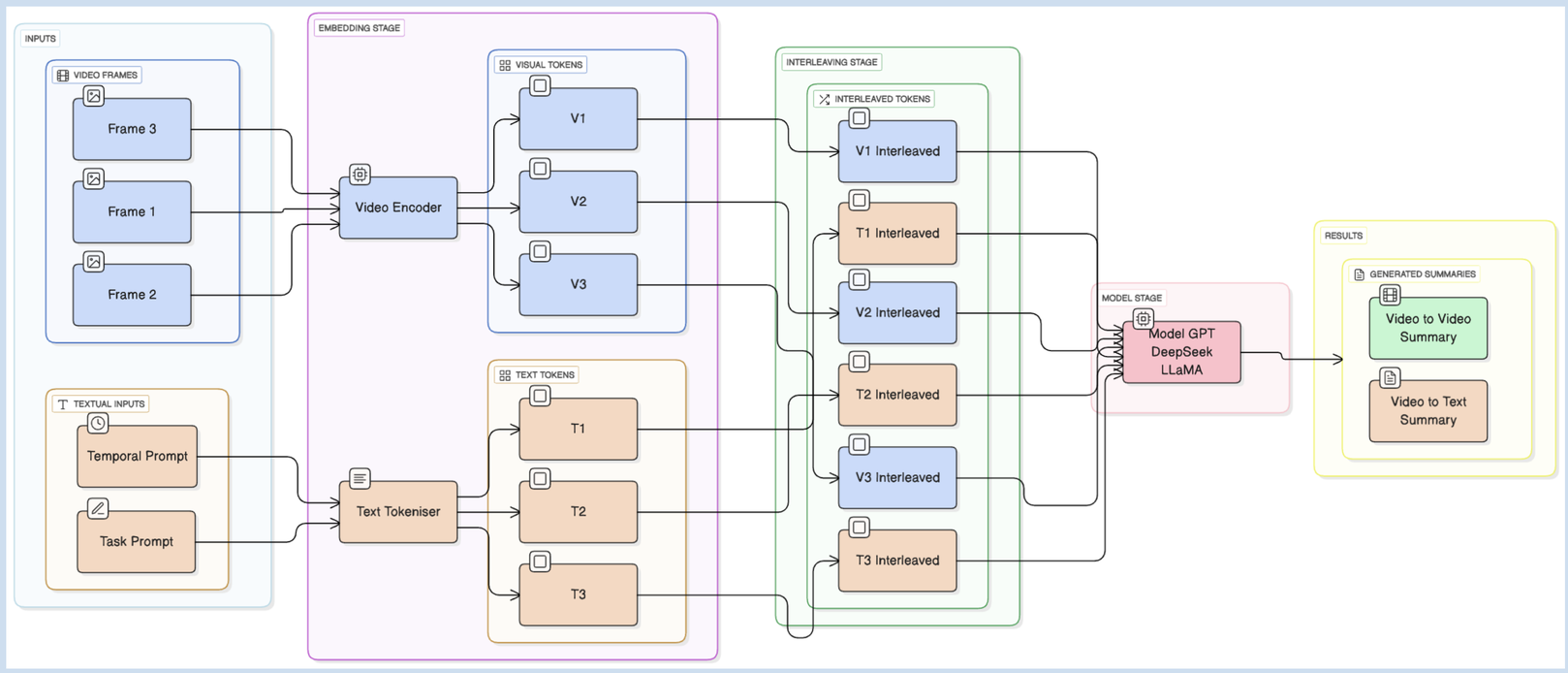

Fig1: The architecture of the proposed technique

Results

Qualitative Example - Prototype of Base Model:
Query: Please generate BOTH video and text summarization for this video.
Answer: A man is seen standing in a room and begins to dance [10, 11, ... 19]. He continues to dance and ends by walking away [89, 90, ... 99].

Conclusion & Future Work

This project presents a comparative and efficient cross-modal summarization pipeline, as LLaMA + CLIP. Next work is to evaluate all results and:

- Incorporate audio and subtitle cues for richer context.

References

[1]Dataset : Activity Net-Cap https://www.kaggle.com/datasets/akshayajit/activitynet-captions