Vision LLM-based Cross-modal Summarization Framework for Long-range Videos

Shicheng Peng, Dingrui Tao, Ruiyang Yu

College of Literature, Science, and the Arts, University of Michigan

Introduction

Video summarization is a type of technique that aims to summarize a long video, producing a concise and comprehensive summary that describes its main idea. As Figure 1 shows, video summaries can have various forms, among which **video and text summaries** are two typical forms.

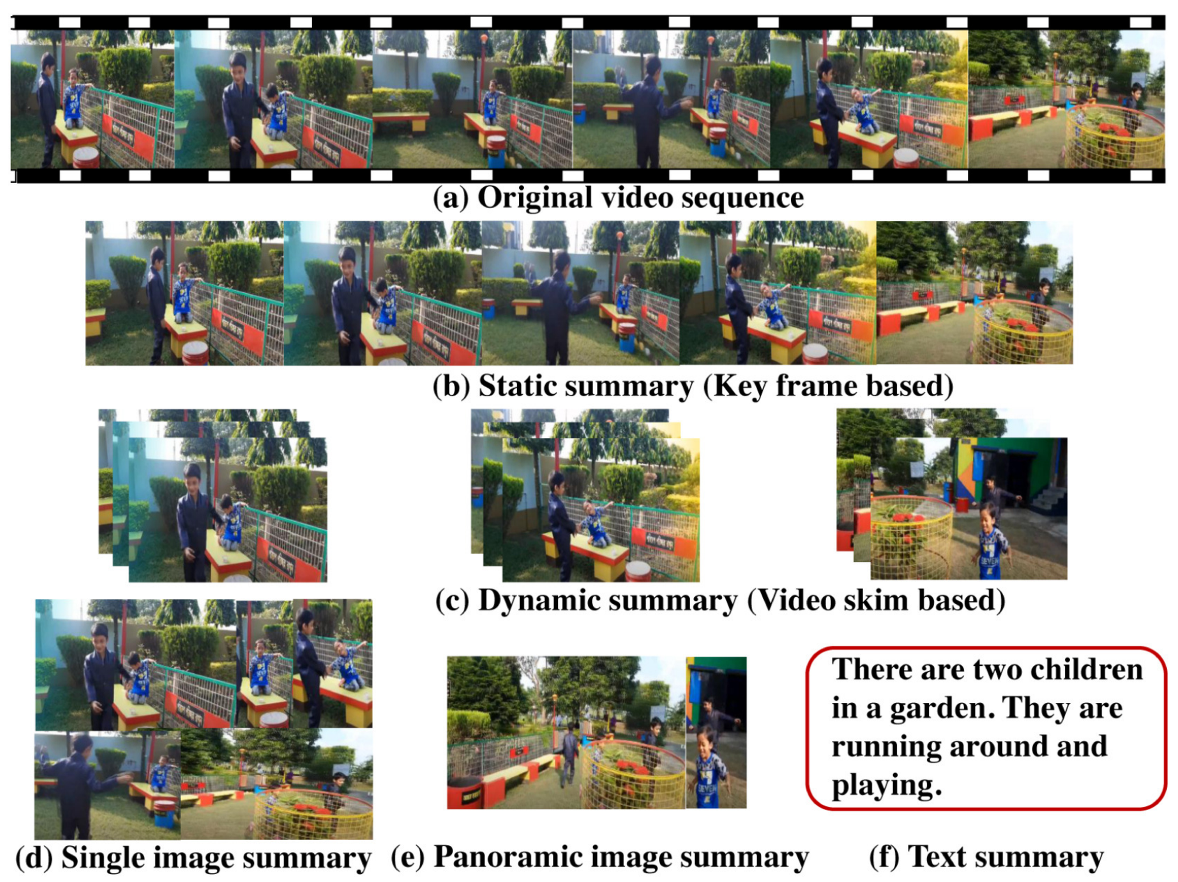


Figure 1. Different types of video summary

However, previous work often treats video and text as independent tasks, ignoring the possible association between them in the videos, especially for long-range versions. Thus based on SOTA vision LLM, we proposed a framework to facilitate the understanding of cross-modal information in long-range videos. We introduce a **local attention module** as well as **fine-tune** the backbone model.

Motivation & Intuition

The most important part we introduced is an **local attention module**. Previous study have shown that it is a effective and powerful method when dealing with multi-modal information in video summarization tasks.

- Local Attention helps model to learn both local and global representations from videos (Lan et al., 2025)
- A Cross-Modal Attention module enables joint reasoning between texts and frames of a video (Gorti et al., 2022)
- Cross-Modal Attention module allows the model to adaptively focus on important video segment and keywords (Ye et al., 2022)
- Local contextual attention can reduce redundancy by extracting representative information in short video segments (Pan et al., 2022)

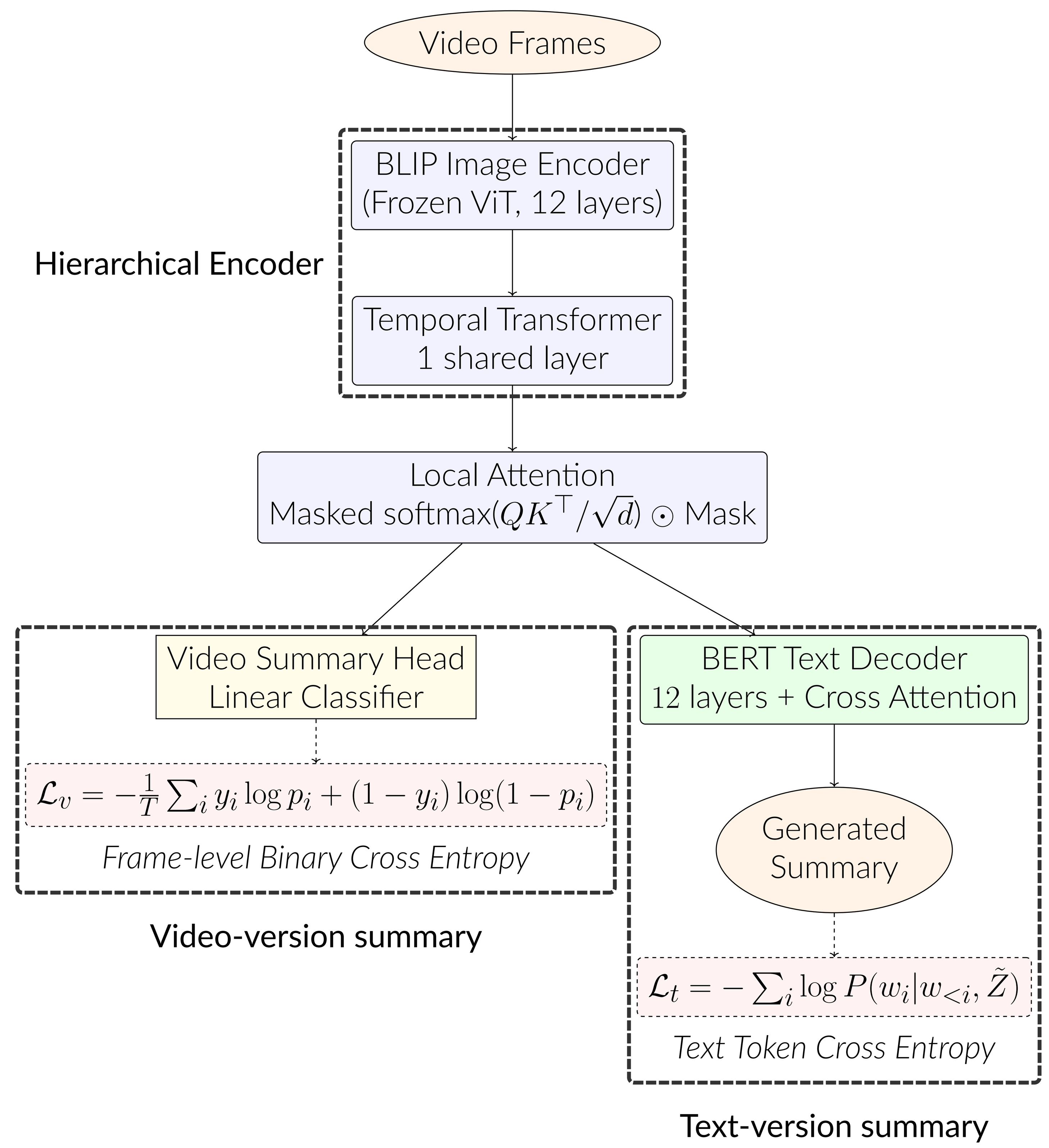
Data Set

VideoXum is a dataset based on **ActivityNet Captions** dataset with human re-annotation, which is perfect for cross-modal learning.

- **14,000+** $10 \sim 755$ **seconds open area** Youtube videos with 1 fps capturing
- **140,000 video-caption pairs** (1 video is annotated with 10 captions)
- Each clip is graded with **saliency label** by human
- **Training (8000), validation (2000) and testing (4000)**

Method

Our framework leverage the large vision-language pre-trained model BLIP as the backbone. **The main structure** contains four major parts: **1. Hierarchical Encoder** for learning representation. **2. Local Attention module** to learn local dependency between video and text content. **3. Video-Summary Decoder** designed for decoding representation into video summary. **4. Text-Summary Decoder** designed for decoding representation into text summary. Details see below:



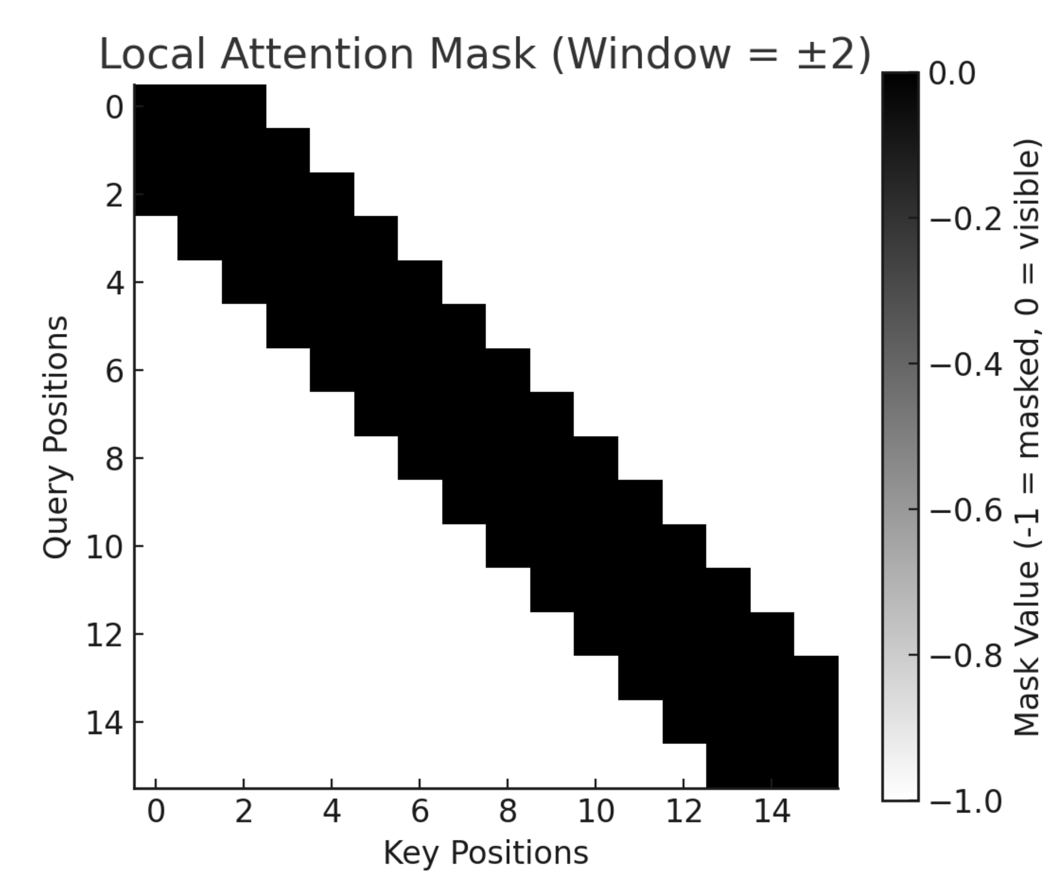


Figure 2. Local Attention Mask with k = 2

Result

Metric Comparison

Table 1. Performance Comparison Across Models with Highlighted Improvements

Metric	Base	TT+CA	TT+CA+LA
F1 (V2V)	21.7	23.5	23.6
Kendall	0.131	0.196	0.201
Spearman	0.207	0.258	0.269
BLEU@4	5.5	5.8	5.9
METEOR	11.7	12.2	12.4
ROUGE-L	24.9	25.1	25.1
CIDEr	18.6	23.1	25.4
VT-CLIPScore	28.4	29.4	29.6

Sample Illustration

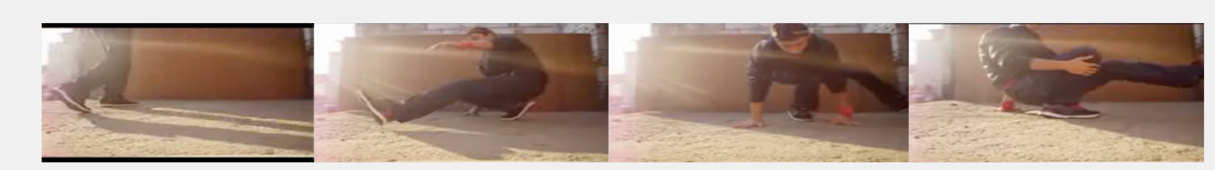


Figure 3. Sample Video

V2V-SUM

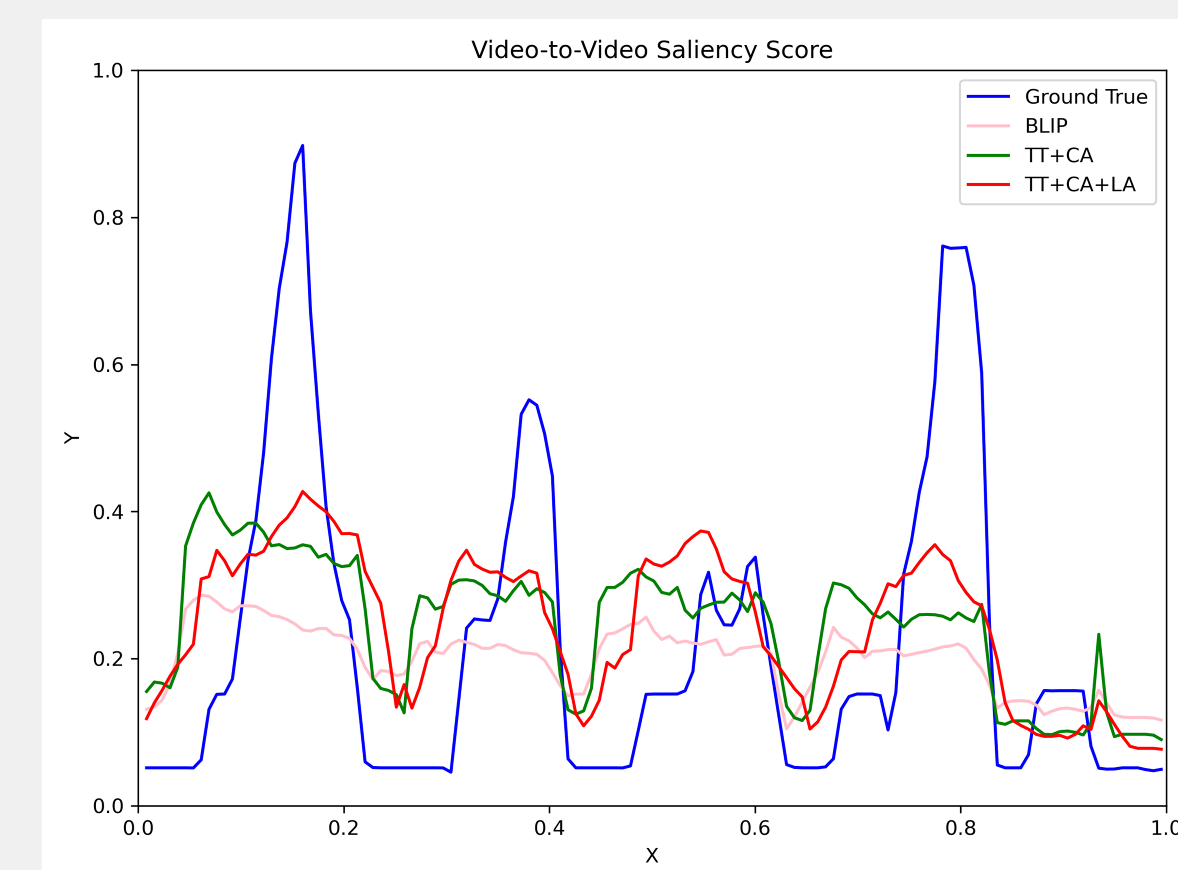


Figure 4. Saliency Score Comparison

V2T-SUM

several shots of text are shown followed by a **person walking into frame** the person then walks in and out of frame performing various **dance moves on the ground** the person **continues moving around** on the ground and demonstrating **how to perform moves**

a man is seen standing in front of a camera and leads into him bending down on the ground the man is then seen bending down on a piece of wood and leads into him bending down on the ground

a close up of a shoe is shown followed by a **person stepping up and down** on the shoe the person **continues stepping up and down** on the shoe while **the camera captures his movements**

a close up of a shoe is shown followed a **person moving around the ground** the person is then **bending down kicking and sweeping** across the ground on the shoe as **the camera tracks his motions**