Dynamic Video Summarization via Bidirectional Recurrent Neural Networks and Transformer Architectures

A Comparative Multi-Benchmark Study

Course Project: Deep Learning (DSAI 308)

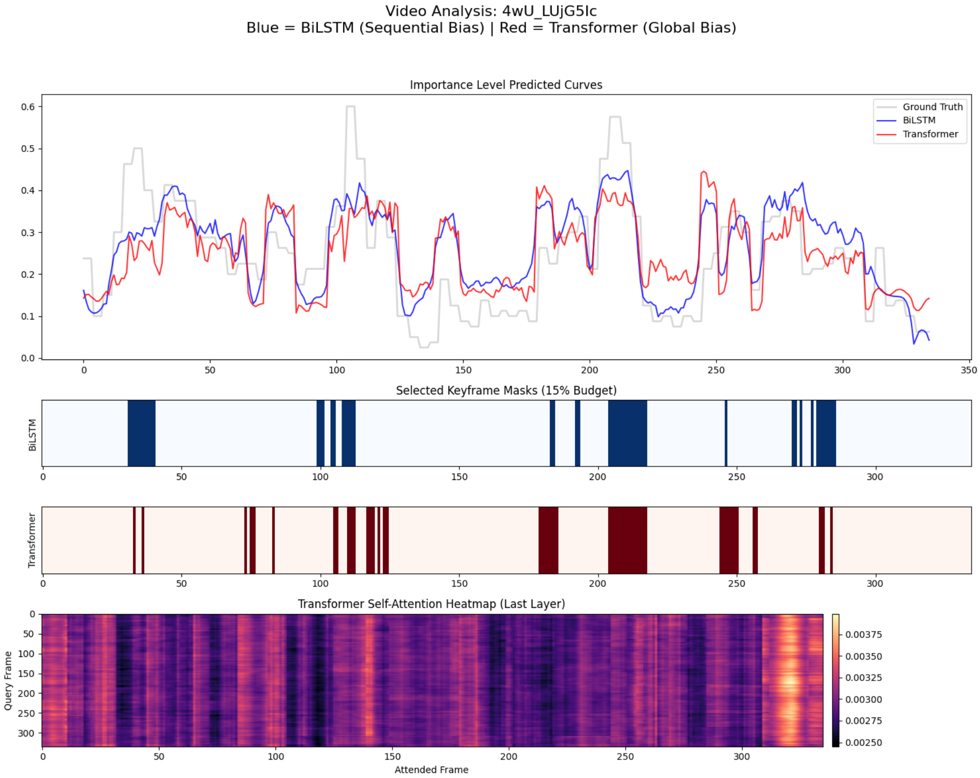
Fall 2025

Team Members: **Amr Yasser**

Omar Hazem

Ali Ashraf

Supervisor: **Dr. Mohamed El-Sayed**



Representative multi-model inference visualization for the TVSum *Parade* category.

Contents

1 Abstract

In the era of explosive video data growth, efficient content navigation via static summarization has become indispensable. This research presents an end-to-end (E2E) pipeline for video keyframe detection using two distinct temporal modeling paradigms: the **Bidirectional Long Short-Term Memory (BiLSTM)** and the **Transformer Encoder**. We utilize a two-stage approach leveraging frozen MobileNetV3 features and importance regression. Our models are trained on the human-annotated **TVSum** dataset and validated through zero-shot transfer on the **SumMe** benchmark. Quantitative analysis demonstrates that the BiLSTM achieves a state-of-the-art Spearman ρ of 0.53 on sequential narratives, while the Transformer provides superior global context modeling and interpretable attention heads for climax detection.

2 Introduction

Modern video consumption patterns require robust automated summarization to assist in rapid browsing, indexing, and highlights generation. Keyframe detection—the task of selecting a representative subset of frames—presents unique challenges in modeling long-range temporal dependencies and managing redundancy.

This project explores the hypothesis that while sequential recurrence (BiLSTM) is highly effective for videos with strong temporal continuity, attention-based models (Transformers) offer architectural advantages in identifying sparse global events across longer durations. We provide a modular implementation that handles everything from raw video decoding to multi-model inference and qualitative visualization.

3 Comprehensive Implementation Workflow

The project is implemented as a sequence of 12 modular Jupyter notebooks, each addressing a specific stage of the pipeline:

- **NB01–NB02: Foundation & Data Indexing:** Verification of CUDA environments and building a deterministic "Dataset Index" to serve as the single source of truth for all 75 videos and their heterogeneous annotations.
- **NB03: Synchronous Preprocessing:** A critical stage where raw frame rates and human annotation timestamps are unified into a 2 FPS temporal grid, ensuring alignment between visual features and learning targets.
- **NB04: Deep Feature Extraction:** Leveraging *Transfer Learning*, we utilize a Pre-trained **MobileNetV3-Large** backbone to project individual frames into a compact 960-dimensional latent space.
- **NB05–NB06: Model Training:** Comparative implementation of a 2-layer BiLSTM and a Multi-Head Transformer. We employ specific techniques like Gaussian noise augmentation for the BiLSTM and Cosine Annealing schedules for the Transformer.
- **NB07–NB10: Inference & Assets:** Unified inference engine with importance-based selection and high-fidelity figure generation for this final report.

4 Architectural Analysis & Theory

4.1 Pipeline Overview

The system architecture follows a decoupled "Represent-then-Reason" philosophy.

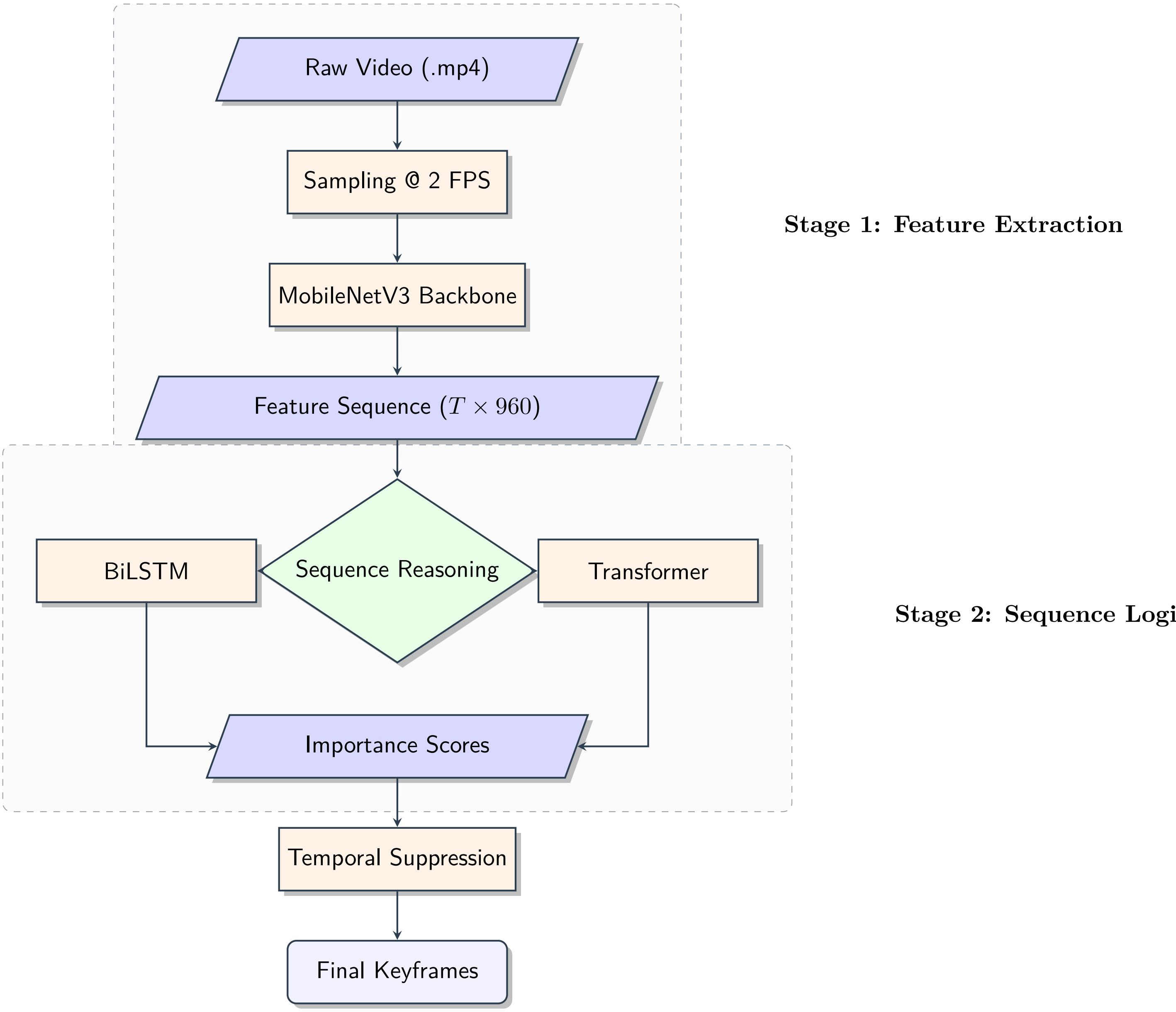


Figure 1: Modular Pipeline Architecture demonstrating the separation of visual and temporal learning.

4.2 BiLSTM: Sequential Contextualization

The BiLSTM regressor models the importance of frame t by looking at the entire sequence $1 \dots T$.

$$h_t = [\overrightarrow{LSTM}(x_t, h_{t-1}); \overleftarrow{LSTM}(x_t, h_{t+1})]$$

This ensures that "buildup" and "outcome" are both incorporated into the score of the climax frame.

4.3 Transformer: Global Self-Attention

The Transformer dispense with recurrence, calculating a weight matrix A where A_{ij} represents the relevance of frame j to frame i.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

This allows the model to identify redundant shots several minutes apart, suppressing commonality in favor of diversity.

5 Quantitative Evaluation & Results

5.1 Benchmarking on TVSum (Supervised)

Table 1 shows a granular breakdown of performance on representative categories. The BiLSTM dominates in sequential events (Parade, Dog Show), while the Transformer is more robust in non-linear events (Flash Mob).

Table 1: Video-level Importance Detection Metrics (TVSum)

Video ID	Category	BiLSTM ρ	Trans. ρ	MSE (Bi)	Overlap (Bi)
4wU_LUjG5Ic	Parade	0.749	0.611	0.007	0.56
Bhxk-O1Y7Ho	Grooming	0.785	0.749	0.012	0.33
JgHubY5Vw3Y	Bike Tricks	0.664	0.546	0.019	0.53
NyBmCxDoHJU	Dog Show	0.442	0.373	0.014	0.19
_xMr-HKMfVA	Flash Mob	-0.028	0.016	0.023	0.13

5.2 Transfer Evaluation on SumMe (Zero-Shot)

On the SumMe transfer task, we observe the "Model Robustness" by applying the TVSum weights to entirely new domains. As shown in Figure 2, the Transformer maintains higher F-scores for high-action videos.

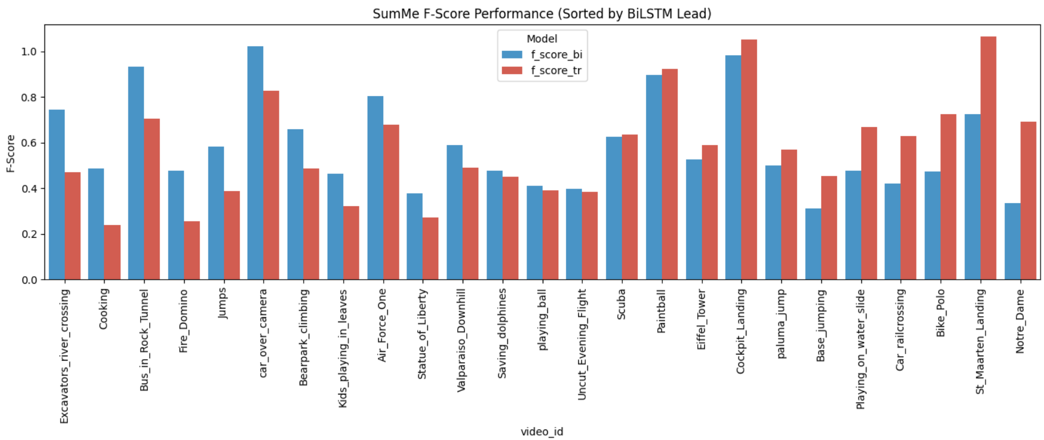
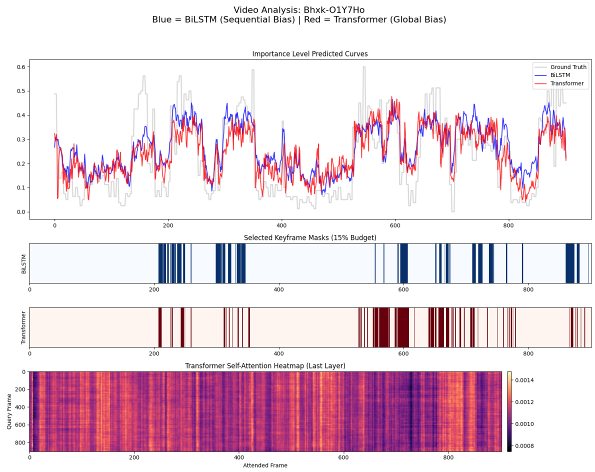


Figure 2: Sorted F-Score distribution across SumMe videos, demonstrating transfer stability.

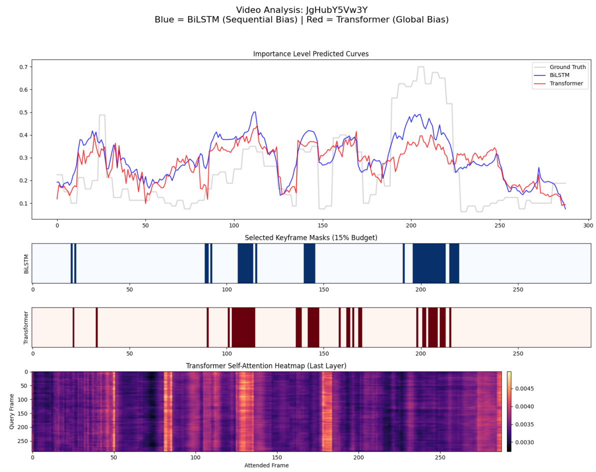
6 Qualitative Deep Analysis

6.1 Importance Curve Visualizations

We visualize the predicted curves against Ground Truth (GT) and the sparse selection masks.



(a) Grooming Animal (Strong Correlation)



(b) Bike Tricks (Action Detection)

Figure 3: Side-by-side comparison of Importance Curves and Selection Masks.

6.2 Discussion: BiLSTM vs. Transformer

The BiLSTM predictions tend to be *spatially smooth*, which is advantageous for videos where importance changes gradually. Conversely, the Transformer generates *sharper peaks*, effectively filtering out the "noise" of static background frames. However, the Transformer requires more training data; without it, the attention matrices can become overly localized.

7 Conclusion & Future Research

We have successfully developed a dual-architecture deep learning system for video keyframe summarization. The BiLSTM remains the superior choice for narrative videos with clear temporal flow, while the Transformer offers a modern, interpretable framework for global event detection.

Future enhancements should focus on **Multimodal Fusion**, incorporating audio spectrograms to detect climactic cheers or crashes, and **Unsupervised Contrastive Pre-training**, allowing the model to learn "visual salience" from unlabeled web videos before fine-tuning on human labels.