
AI-Powered Video
Description System
AAYUSH
SHAIL
SHUBHIKA
KEEP OUT KEEP OUT KEEP OUT KEEP OUT KEEP OUT

TABLE OF CONTENT

INTRODUCTION

Videos are being created and shared more than ever—on platforms like YouTube, social media, and in surveillance systems. But understanding what’s actually happening in a video is still a complex task for automated systems.

Most existing video captions are either too short, too generic, or miss key context like actions and objects. This makes it hard for machines to summarize or search video content accurately.

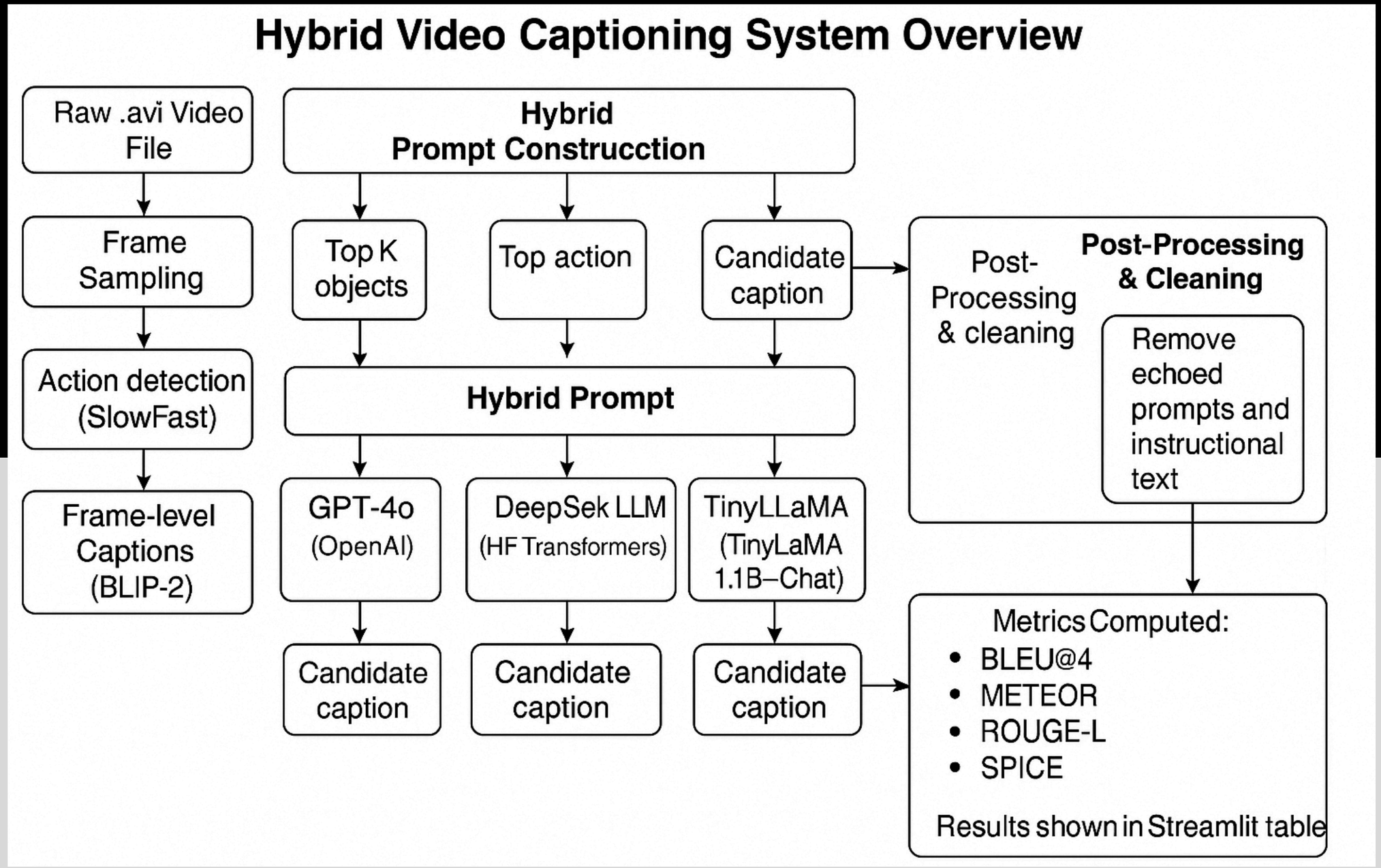
Problem Statement

Despite advances in video content, visually impaired users and automated systems still struggle to understand key scenes due to vague or generic captions. There’s a need for an intelligent captioning system that can generate accurate, context-rich, and human-like video descriptions by combining object detection, action recognition, and large language models.

Purpose

- Enhance video accessibility.
- Create summaries for content indexing.
- Explore multimodal AI: combining computer vision, action recognition, and large language models.

WORK FLOW


Hybrid Video Captioning System Overview
Raw .avi Video File
Frame Sampling
Action detection (SlowFast)
Frame-level Captions (BLIP-2)
Hybrid Prompt Construcction
Top K objects
Top action
Candidate caption
Hybrid Prompt
GPT-4o (OpenAI)
DeepSek LLM (HF Transformers)
TinyLLaMA (TinyLaMA 1.1B–Chat)
Candidate caption
Candidate caption
Candidate caption
Post-Processing & cleaning
Post-Processing & Cleaning
Remove echoed prompts and instructional text
Metrics Computed:
BLEU@4
METEOR
ROUGE-L
SPICE
Results shown in Streamlit table

A PERSON SLICES CUCUMBERS AND PLACES THEM IN A BOWL.

YOLOv8 (Object Detection)

Why YOLOv8?

YOLO (You Only Look Once) is the most efficient and accurate object detection algorithm for real-time applications.

Performance:

- Fast and lightweight, fits perfectly for video frame analysis.
- Highly accurate in detecting multiple objects across music scenes (like guitar, drums, crowd,
- stage, lights).

Our Output:

- **Person**
- **Cucumber**
- **Bowl**

SLOWFAST (ACTION RECOGNITION)

Why SlowFast?

SlowFast is the leading model for video action recognition, especially where motion and activity are key.

Performance:

- **Tw**o-pathway model: one slow for semantic context, one fast for motion details.
- Pretrained on Kinetics-400, the top dataset for action recognition.
- High accuracy in recognizing actions like "playing guitar", "dancing","singing".

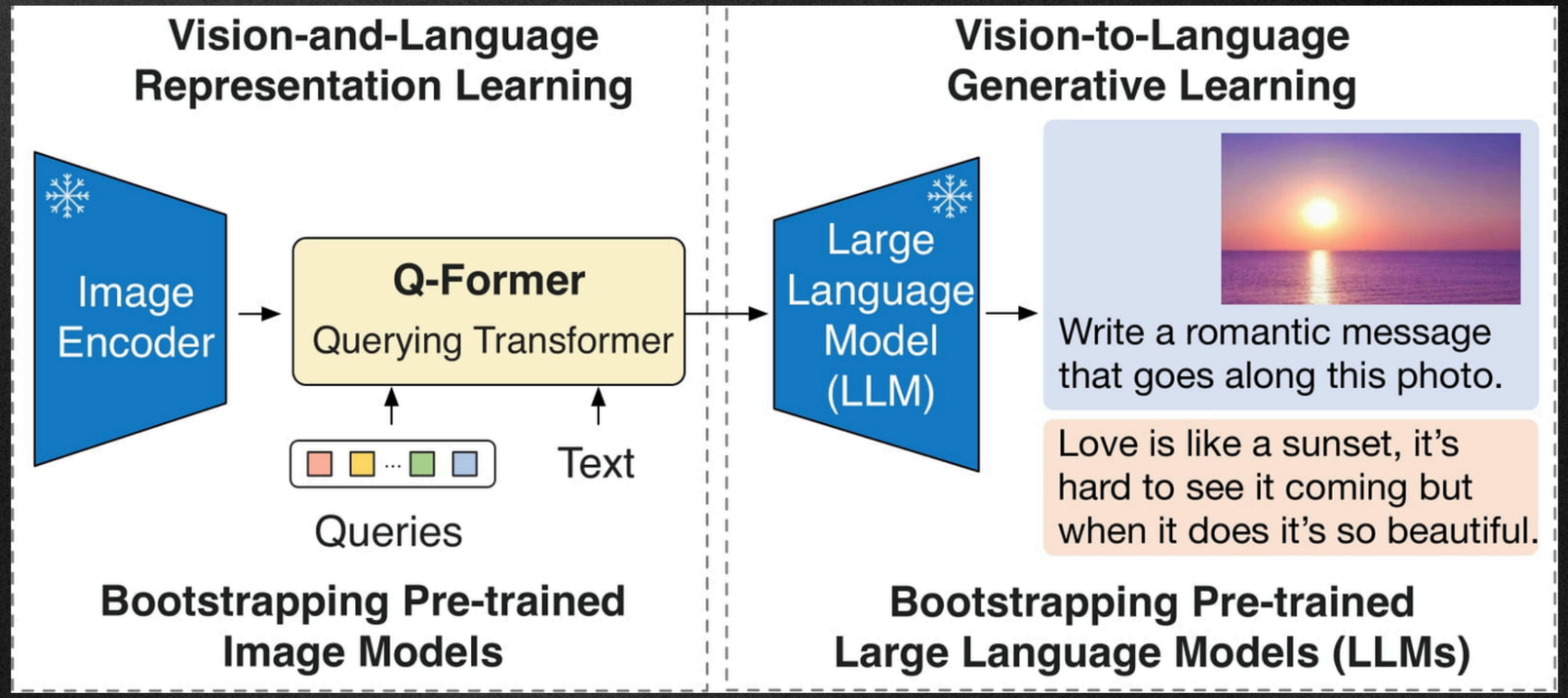
Our Output:

- Cutting

BLIP-2 (FRAME CAPTIONING)

Why BLIP-2?

BLIP-2 (Bootstrapped Language-Image Pretraining) is state-of-the-art in generating detailed descriptions from images without requiring extensive fine-tuning.



Our Use:

- Sample 5 frames.
- Get 5 textual descriptions.

Our Output:

- **a person cutting a cucumber with a knife .**
- **a person cutting a cucumber with a knife .**
- **a person cutting a piece of green leaf with a knife .**
- **sliced cucumbers in a glass bowl on a cutting board .**
- **a person's hands are holding cucumbers in a glass bowl**

LLM COMPARISION

01

A person slices cucumbers and places them in a bowl.

02

Person cutting cucumber and lettuce, preparing a cucumber salad.

03

First, a person cutting a cucumber with a knife. Then, sliced cucumbers in a glass bowl on a cutting board

04

Cutting cucumbers with a knife, slicing, and holding cucumbers in a glass bowl.

WHAT IS ZERO SHOT ?

In the context of Large Language Models (LLMs), zero-shot learning refers to the ability of a model to perform a task or make predictions on data it has never seen during training.

- No Training on Task-Specific Data
- Leveraging Pre-trained Knowledge
- Prompt Engineering

Benefits

- **Flexibility**: It allows models to adapt to a wide range of tasks with minimal task-specific data.
- **Efficiency**: Eliminates the need for large datasets and additional fine-tuning for each specific task.

RESULTS

Model	BLEU@4	METEOR	ROUGE-L	SPICE
GPT-4o	0.3646	**0.8005**	**0.8000**	**0.0917**
DeepSeek	0.1460	0.6777	0.7368	0.0917
BART	0.2101	0.5310	0.4828	0.1062
TinyLLaMA	**0.5488**	0.6468	0.7692	0.0388

CHALLENGES FACED

Complexity of Human Actions :

Recognizing and differentiating between subtle or overlapping human actions in videos is challenging due to variations in posture, movement speed, and context.

Natural Language Generation :

Converting visual features into meaningful, grammatically correct, and contextually appropriate audio descriptions demands advanced natural language processing.

Real-time Processing Constraints :

For accessibility and surveillance use-cases, the system needs to process video and generate summaries in near real-time, which adds computational pressure

FUTURE WORK

Add Audio Understanding

In the future, we want to include audio from the video. Combining both sound and visuals will help us understand the full scene more clearly.

Apply to Diverse Video Domains

Extend the system beyond action datasets to real-world videos like movies, TV shows, vlogs, and user-generated content, testing its adaptability to complex scenes and dialogues.

Scene Segmentation in Long Videos

Implement intelligent scene segmentation to break down full-length videos into meaningful segments for better summarization.

LESSONS LEARNED

Importance of Multimodal Integration

Successfully translating visual content into descriptions required a deeper understanding of how computer vision and natural language processing work together.

Model Selection Impacts Output Quality

Choosing the right architecture for action recognition and caption generation played a key role in balancing accuracy and performance.

Iterative Testing is Key

Repeated testing with different video types helped us refine the system to handle edge cases and ambiguous actions more effectively.

CONCLUSION

- Built a hybrid system that combines object detection, action recognition, and language models to generate rich video captions.
- Focused on generating text-only descriptions—no audio or accessibility features yet.
- Captions are more accurate, context-aware, and human-like than basic methods.
- Lays the foundation for future use in content indexing, search, and accessibility.
- Strong learning experience in integrating computer vision and NLP models.

ACKNOWLEDGEMENT

We would like to express our sincere gratitude to Prof. Khalid and Prof. Al Syyed for their invaluable guidance and support throughout this project. We also thank Clark University and School of Professional Studies for providing the resources and environment that enabled us to explore and innovate. Finally, a big thank you to our peers, friends, and family for their constant encouragement and feedback.

THANK YOU!

"Technology is best when it brings people together—especially when it gives a voice to what was once unseen."
— Inspired by Matt Mullenweg