NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

School of Computer Science and Engineering
College of Engineering

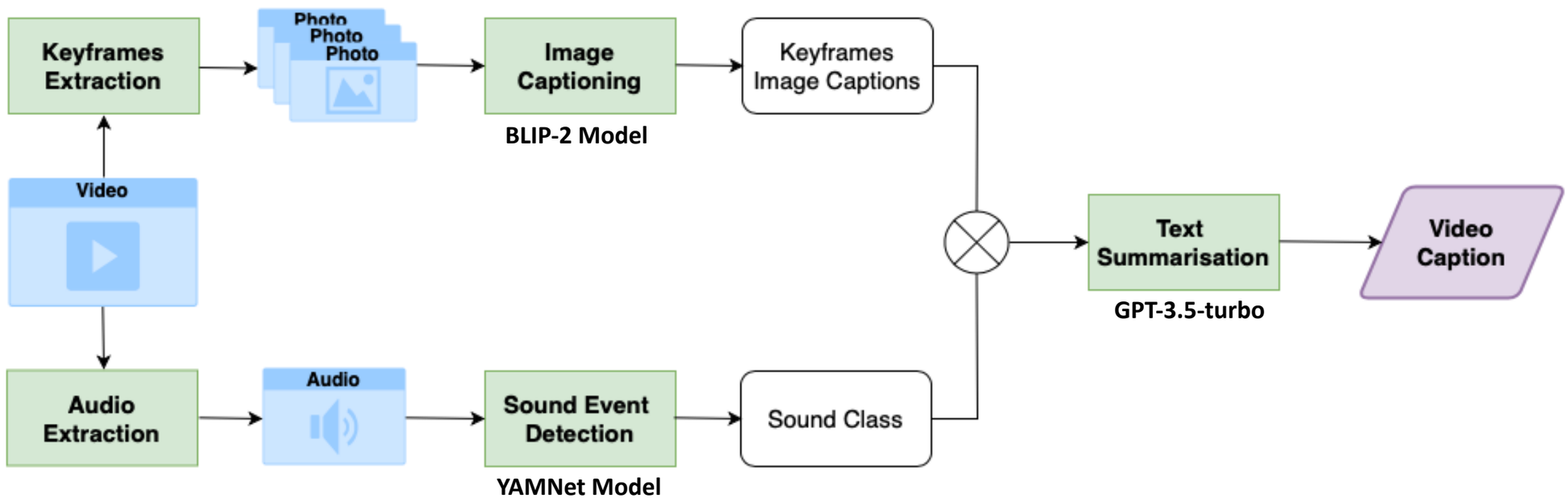
Neural Image & Video Captioning

Student: Lam Ting En Supervisor: A/P Zhang Hanwang

Project Objectives:

This project aims to develop a video captioning model capable of generating multimodal captions from video content. Extending from the state-of-the-art image captioning model BLIP-2, the video captioning model integrates keyframes extraction, image captioning, sound event detection and text summarisation.

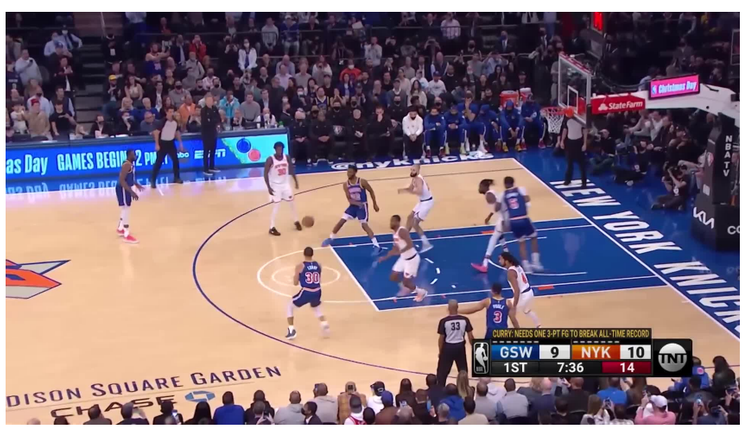
System Architecture:



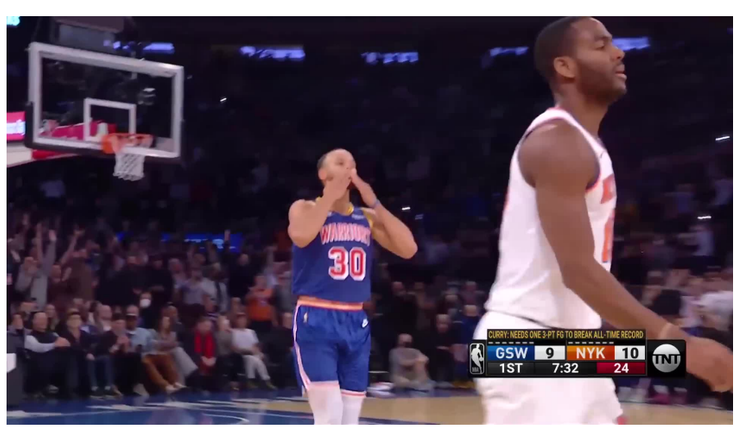
Example:

A clip of an action-packed basketball game is used for demonstration. The clip has a duration of 8 seconds and depicts a player scoring a basketball goal, the crowd cheering and the player celebrating.

Keyframes Extraction & Image Captioning



frame-0001: a basketball game is being played in front of a crowd.



frame-0005: stephen curry's game-winning shot vs knicks

frame-0007: a basketball player is celebrating in front of a crowd.

Sound Event Detection

```
Top 3 inferred classes and their scores:
Crowd : 0.2035
Speech : 0.1914
Cheering : 0.0633
```

Text Summarisation → Final Video Caption

'The video captures a basketball game with a crowd in the stands. The highlight is Stephen Curry's game-winning shot against the Knicks, leading to a celebration by players and fans cheering and speech heard in the background.'

Evaluation:

The model-generated captions were evaluated by comparing them with the annotations of the MSVD and MSR-VTT datasets.

Dataset	Evaluation Metrics			
	BLEU-4	ROUGE-L	METEOR	CIDEr
MSVD	0.079	0.322	0.209	0.015
MSR-VTT	0.071	0.290	0.200	0.017

A human evaluation study was also conducted with 50 participants with 50 videos from MSVD and MSR-VTT. The selection rate for the model-generated captions was **73.6%** when compared against human annotated captions.