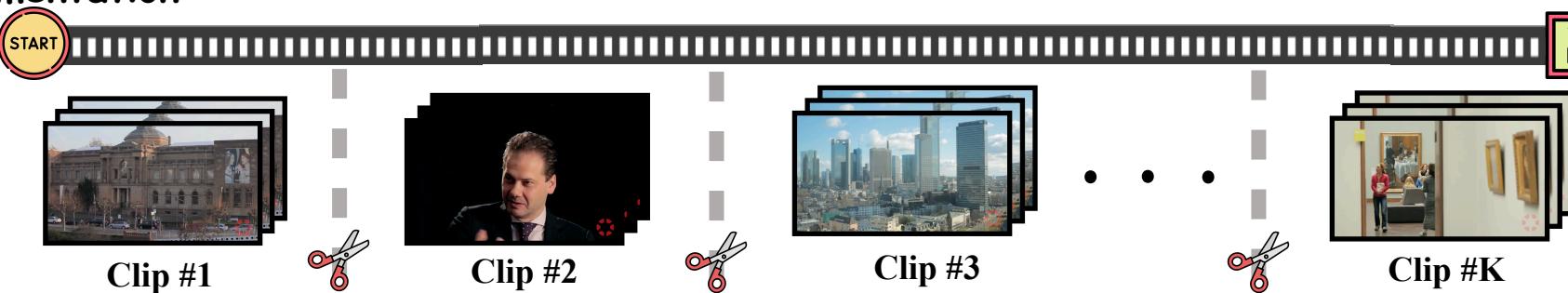


## (a) Video Clip Segmentation



Long Video



## (b) Visual Captioning



Qwen2-VL

Objective Description

From 00:00:00.000 to 00:00:31.798:  
The video opens with a black screen featuring a red circular logo with a white star-like symbol and the text \"mOppen\" in the bottom right corner...

From 00:00:31.832 to 00:00:36.803:  
The video shows a panoramic view of a modern cityscape under a clear blue sky with scattered clouds...

## (c) Audio Captioning

Qwen2-Audio



Video Clip

Narration Environment  
Timbre Event Emotion

From 00:00:00.000 to 00:00:31.798:  
Instrumental music primarily on a keyboard...creates a **dreamy, emotional, and atmospheric** setting...

From 00:00:31.832 to 00:00:36.803:  
A male voice speaks in standard, neutral English...His voice is **clear**, with a **mid-range pitch** and **steady pace**.

## (d) Audio-Visual Description Fusion



Visual Caption



Audio Caption

Qwen2.5

From 00:00:00.000 to 00:00:31.798: ...The music, which is dreamy and atmospheric, sets a contemplative and museum-like ambiance...Visually, the content starts with a black screen displaying the \"mOppen\" logo, a red circle with a white star-like symbol and accompanying text...

From 00:00:31.832 to 00:00:36.803: The scene opens with a panoramic view of a modern cityscape under a clear blue sky dotted with scattered clouds...a male voice begins speaking in standard, neutral English, his tone clear and confident...

## (e) Audio-Visual Question-Answer Pair Generation

Consistency Check  
&  
Redundancy Check



Audio-Visual Caption

Clip Timestamps

Qwen2.5

LongAV-100K

Audio-Visual QA

Audio-Visual-Temporal QA