Freezeing

Training/Fine-tuning

Agent: The video is about a cat and ..., with a ... music

User:
<visual-tokens>\n
<audio tokens>\n
Can you describible the video and the audio?

Large Language Model (Vicuna 7B)

Visual Tokens

Audio Tokens

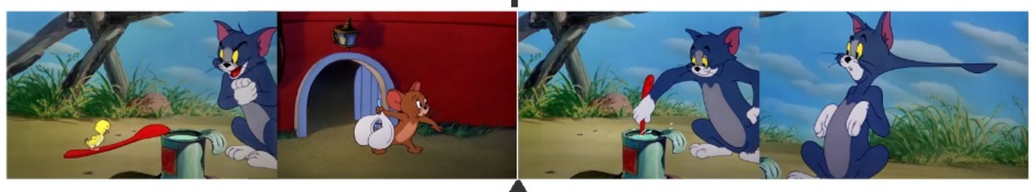
Visual Projection Layer

Audio Projection Layer

Spatial Token Merging

Visual Encoder (CLIP)

Audio Encoder (MERT)



Event-based sampling

Uniform sampling

Input Video