

# Act-ChatGPT: Introducing Action Features into Multi-Modal Large Language Models for Video Understanding

Yuto Nakamizo, Keiji Yanai

The University of Electro-Communications, Tokyo, Japan



## Introduction

- Multi-modal Large Language Models (MLLMs), which integrate visual encoders with Large Language Models (LLMs), have enabled advanced interactive video understanding.
- However, existing MLLMs overlook detailed actions within video segments.

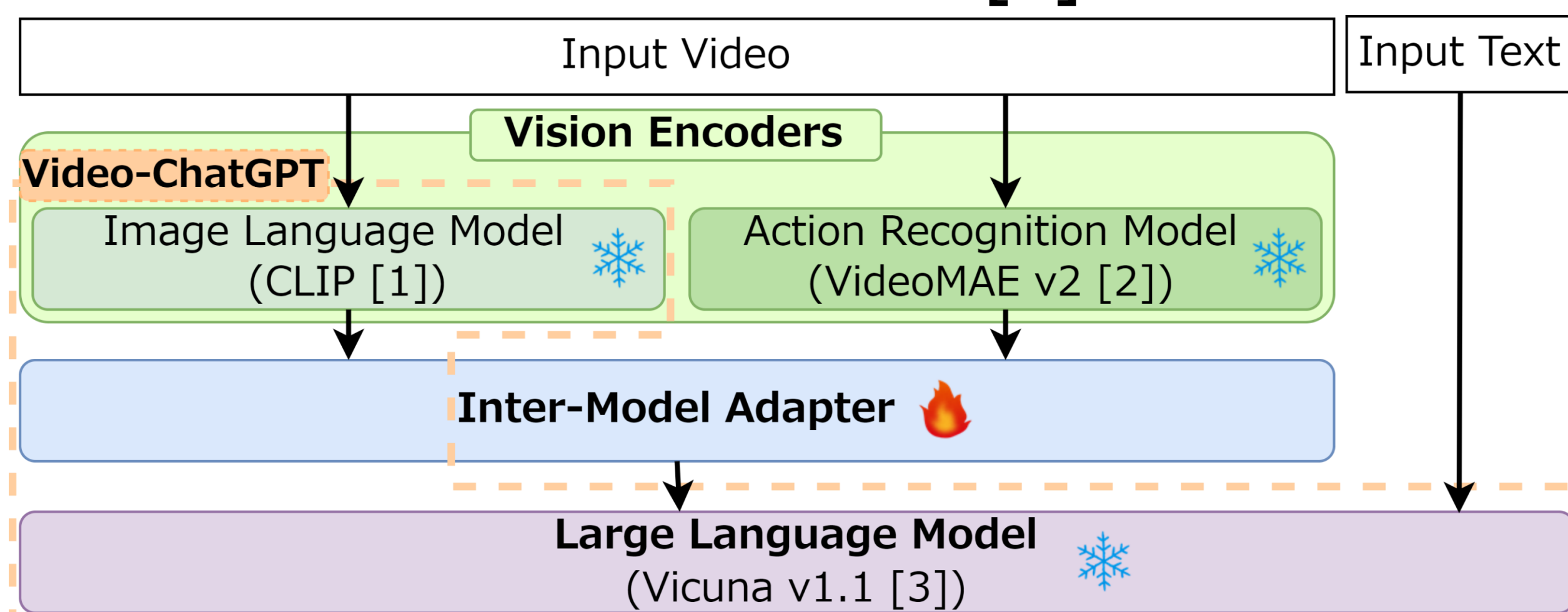
### Objective

For improving action recognition, we introduce action features into the existing MLLM.

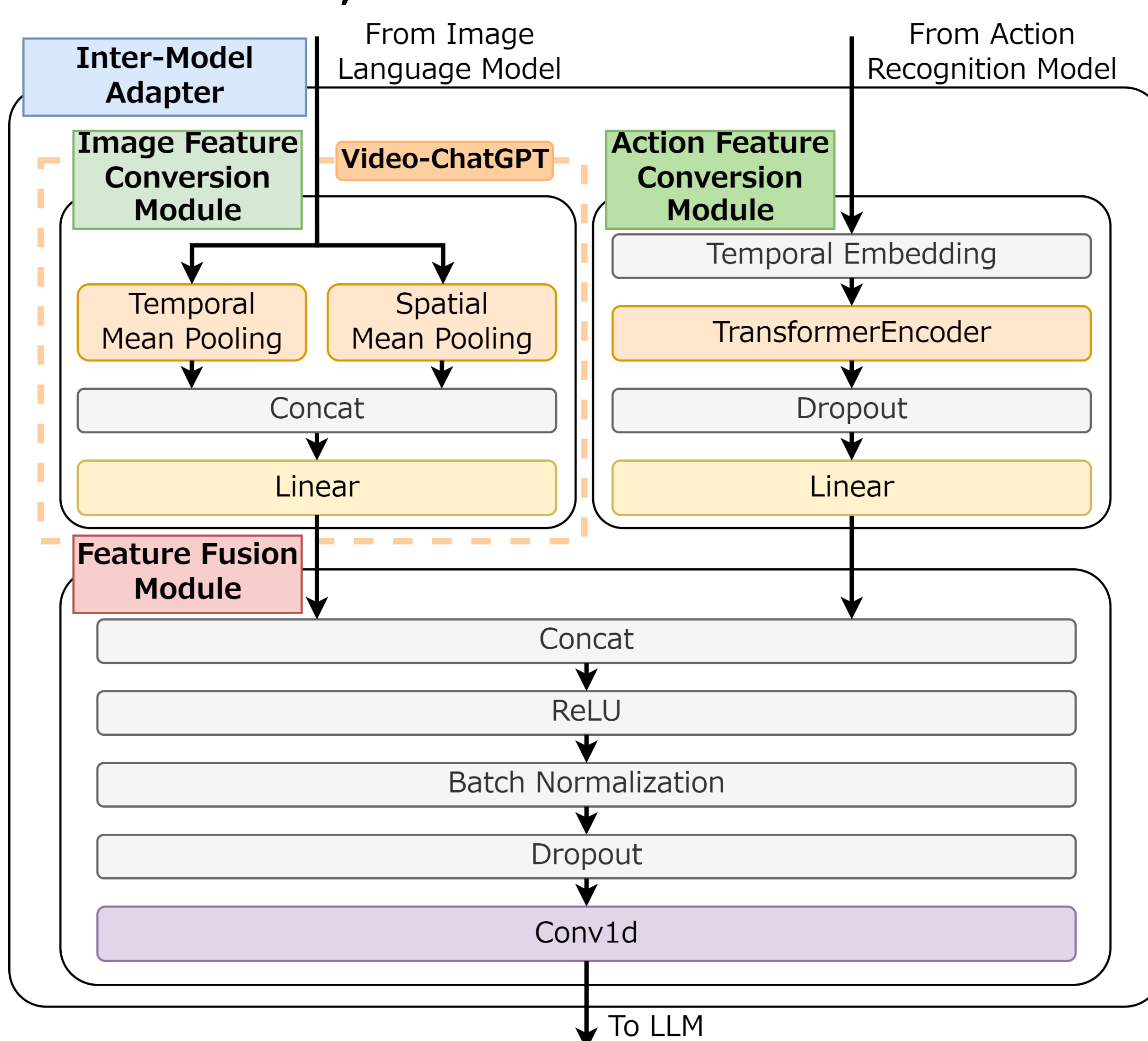
## Method

### Model Architecture

The Proposed method is build by integrating action functions based on Video-ChatGPT [4].



The Inter-Model Adapter is structured from three modules: the Image Feature Conversion Module, the Action Feature Conversion Module, and the Features Fusion Module.



### Training

The training process is conducted in two stages.

- First stage : Training each feature conversion module independently
- Second stage : Training the entire Inter-Model Adapter

### Incorporated data augmentation techniques

- Rephrased existing instruction response texts by LLM

### Training Cost

Training Environments: NVIDIA A5000 48G x 4

- First stage: 15 hours per feature conversion module
- Second stage: 18 hours

## Experiments

### 1. Evaluations on Video-based Generative Performance Benchmarking [4]

Outperformed Video-ChatGPT across all metrics.

手法	CI(↑)	D(↑)	CU(↑)	T(↑)	CO(↑)
Video-ChatGPT	2.50	2.31	2.87	2.10	2.20
Act-ChatGPT (w/o data aug.)	2.53	2.33	2.92	2.13	2.17
<b>Act-ChatGPT</b>	<b>2.62</b>	<b>2.37</b>	<b>3.00</b>	<b>2.20</b>	<b>2.28</b>

※ The evaluation items are Correctness of Information (CI), Detail Orientation (DO), Contextual Understanding (CU), Temporal Understanding (TU) and Consistency (C)

### 2. Evaluations on AutoEval-Video [5]

Outperformed Video-ChatGPT when training from scratch.

手法	Accuracy(↑)
Video-ChatGPT	<b>0.101</b>
Act-ChatGPT	0.064
Video-ChatGPT (scratch)	0.045
Act-ChatGPT (scratch)	0.049

※ (scratch) is trained without initialization of the inter-model adapter by other models.

### 3. Qualitative Evaluations

Improved recognition of actions and related objects while retaining the ability to recognize unique objects.

**Example 1:** Question: "Where is this video taken? What place/landmark is shown in the video?"  
 Video-ChatGPT: "The video is taken at the Statue of Liberty in New York City."  
 Act-ChatGPT: "The video is taken in New York City, specifically at the Statue of Liberty."

**Example 2:** Question: "What unintentional happenings occurred immediately after the man swung the golf club?"  
 Video-ChatGPT: "After the man swung the golf club, he accidentally hit a golf ball into a hole in the sand."  
 Act-ChatGPT: "Yes, the man swung the golf club and immediately fell to the ground."

**Example 3:** Question: "Can you describe the video in detail?"  
 Video-ChatGPT: "Yes, the video shows a young boy performing a handstand on a balance beam. He is seen walking on the beam and then performing a handstand on the beam."  
 Act-ChatGPT: "The video shows a young boy performing gymnastics on a set of parallel bars. He is able to balance on the bars with ease and skill. The video is shot in a gym setting, and the boy appears to be practicing his gymnastics skills."

## Conclusions

We proposed Act-ChatGPT that is MLLM for video understanding using action feature per video segments.

- Introduced action recognition model as an additional visual encoder.
- Outperformed Video-ChatGPT in Video-based Generative Performance Benchmarking.
- Improved recognition of actions and related objects.

[1] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." ICML. PMLR, 2021.

[2] Wang, Limin, et al. "VideoMAE v2: Scaling video masked autoencoders with dual masking." CVPR 2023.

[3] Chiang, Wei-Lin, et al. "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%\* ChatGPT Quality." <https://lmsys.org/blog/2023-03-30-vicuna/>, 2023.

[4] Maaz, Muhammad, et al. "Video-ChatGPT: Towards detailed video understanding via large vision and language models." ACL 2023.

[5] Chen, Xiuyuan, et al. "AutoEval-video: An automatic benchmark for assessing large vision language models in open-ended video question answering." CVPR 2024.