Video-ChatGPT

- □□□□□Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models
- □□□□□https://arxiv.org/abs/2306.05424
- ACL 2024

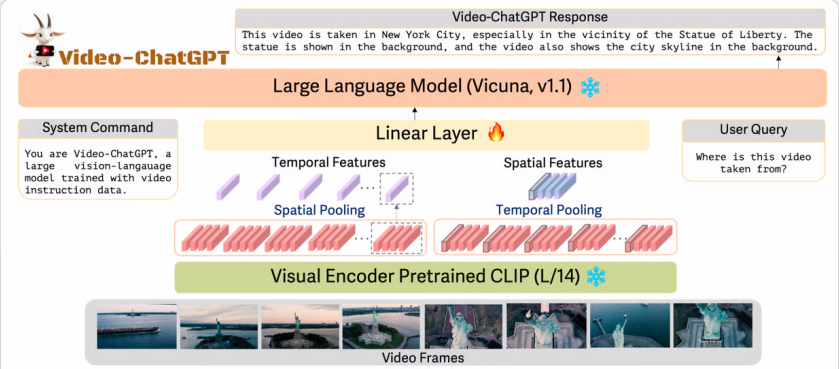

Figure 1: **Architecture of Video-ChatGPT.** Video-ChatGPT leverages the CLIP-L/14 visual encoder to extract both spatial and temporal video features. This is accomplished by averaging frame-level features across temporal and spatial dimensions respectively. The computed spatiotemporal features are then fed into a learnable linear layer, which projects them into the LLMs input space. In our approach, we utilize the Vicuna-v1.1 model, comprised of 7B parameters, and initialize it with weights from LLaVA (Liu et al., 2023).

□□□□