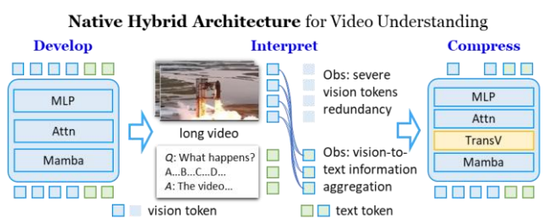
TimeViper:

A Hybrid Mamba-Transformer Vision-Language Model for Efficient Long Video Understanding

Boshen Xu[1], Zihan Xiao[1], Jiaze Li[2], Jianzhong Ju[2], Zhenbo Luo[2], Jian Luan[2], Qin Jin[1]
AIM3 Lab, Renmin University of China[1], MiLM Plus, Xiaomi Inc.[2]

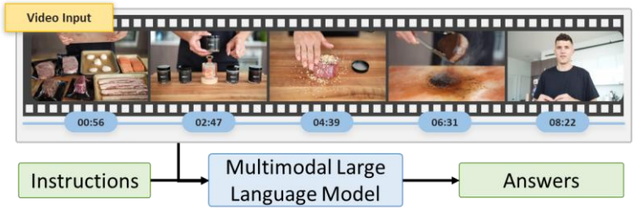


CVPR JUNE 3-7, 2026 DENVER COLORADO

Background

➢ Long video understanding is essential for…

➢ **MLLM paradigm** for hour-long video understanding

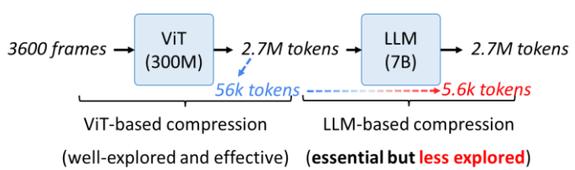


Challenges

➢ **Efficient backbone construction**

Attention	Computation: $O(N^2d)$	$\boldsymbol{O} = \text{softmax}(\boldsymbol{QK^T})\boldsymbol{V} \in \mathbb{R}^{N\times d}$
	KV-Cache: $O(Nd)$	
Linearized or Mamba	Computation: $O(Nd^2)$	$\boldsymbol{S_t} = \boldsymbol{\lambda_t S_{t-1}} + \boldsymbol{v_t k_t^T} \in \mathbb{R}^{d\times d}$
	1 Cache: $O(1d^2)$	$\boldsymbol{o_t} = \boldsymbol{S_t q_t} \in \mathbb{R}^d$

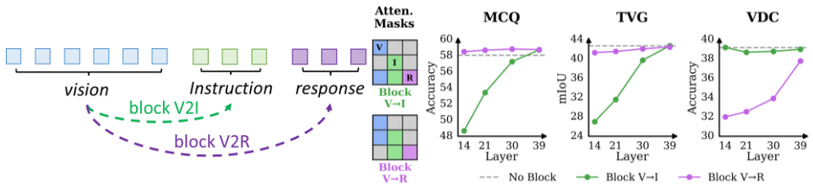
➢ **Severe token redundancy**



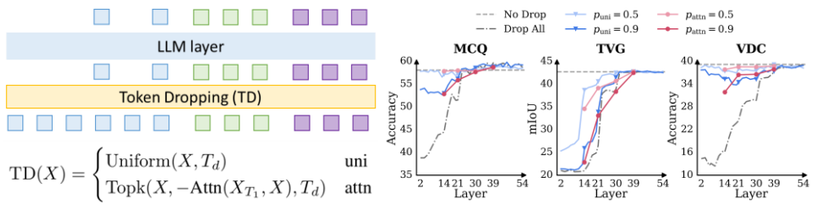
Develop & Interpret & Compress

We first develop a Hybrid Mamba-Transformer VLM baseline to observe…

➢ **Obs1: Vision-to-text information aggregation**
apply attention mask at certain layer for vision token flow

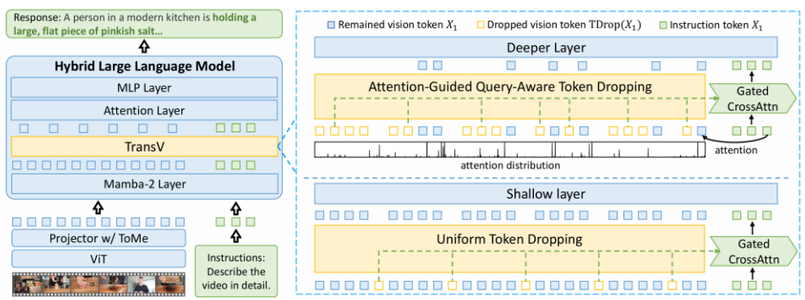


➢ **Obs2: Severe vision tokens redundancy**
apply token dropping with different strategy at certain layer



$$\text{TD}(X) = \begin{cases} \text{Uniform}(X, T_d) & \text{uni} \\ \text{Topk}(X, -\text{Attn}(X_{T_1}, X), T_d) & \text{attn} \end{cases}$$

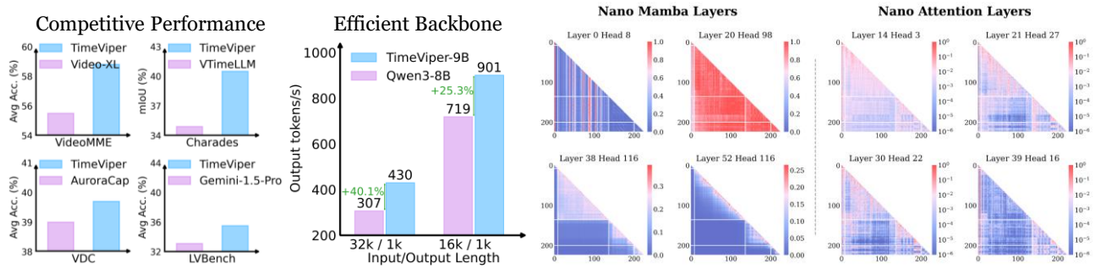
➢ **TransV: Compress a hybrid VLM within LLM**
Dropping rate: 7th shallow layer 50%, 39th layer 90%



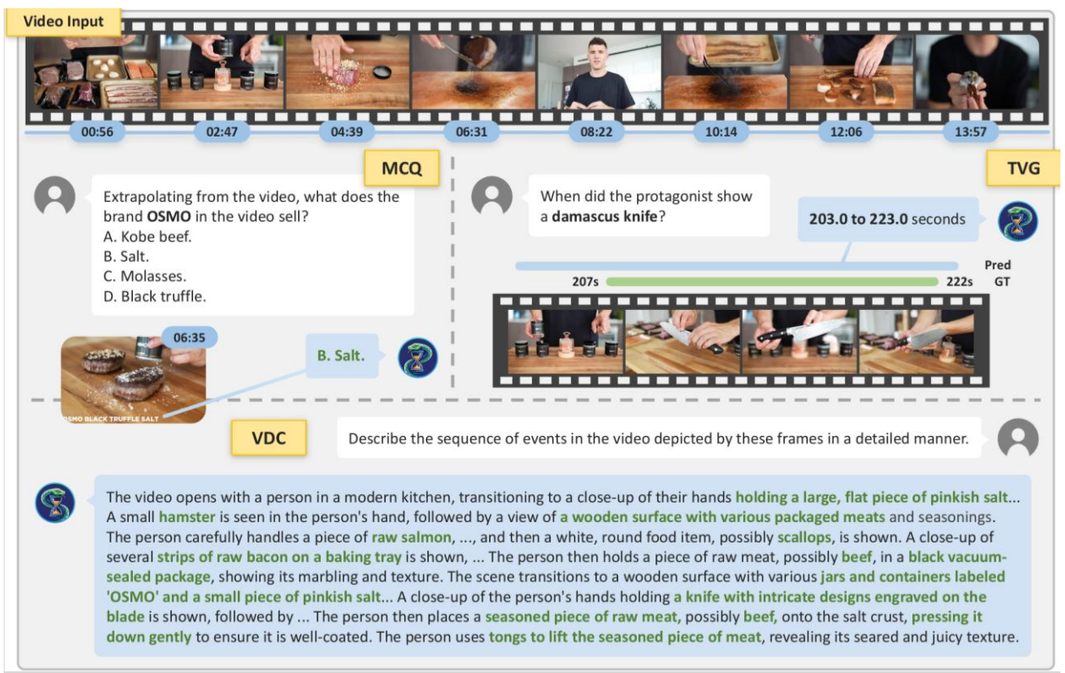
Turn out to be similar to Transformer-based VL processing

Results

➢ Comparisons of benchmark performance, decoding efficiency and attention visualizations



➢ TimeViper works well in video question answering, temporal video grounding and detailed descriptions



More quantitative, efficiency comparison, and interesting visualizations can be found in our paper :)
& Some interesting concurrent works from industry: Nano-VL-9B (NVIDIA), MiMo-V2-Omni (Xiaomi)