STANFORD VISION & LEARNING

Streaming Detection of Queried Event Start

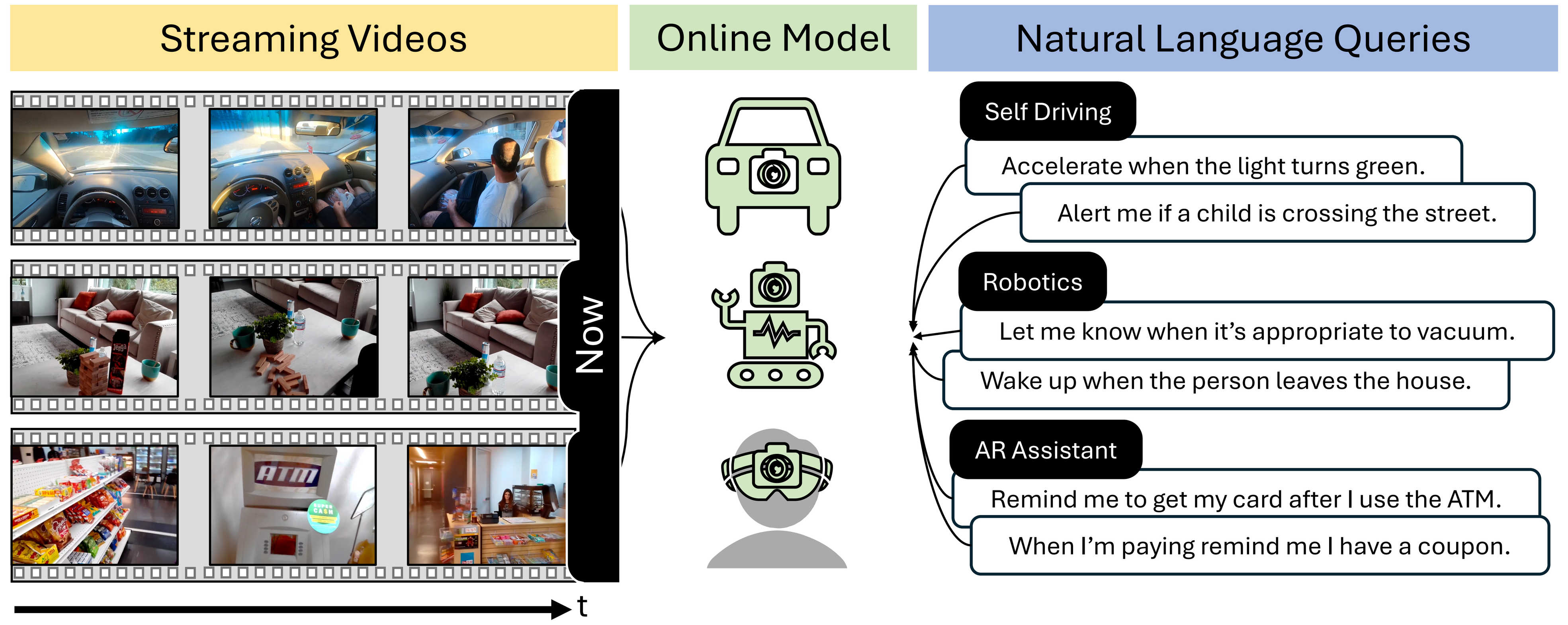
Cristóbal Eyzaguirre, Eric Tang, Shyamal Buch, Adrien Gaidon, Jiajun Wu, Juan Carlos Niebles

NEURAL INFORMATION PROCESSING SYSTEMS

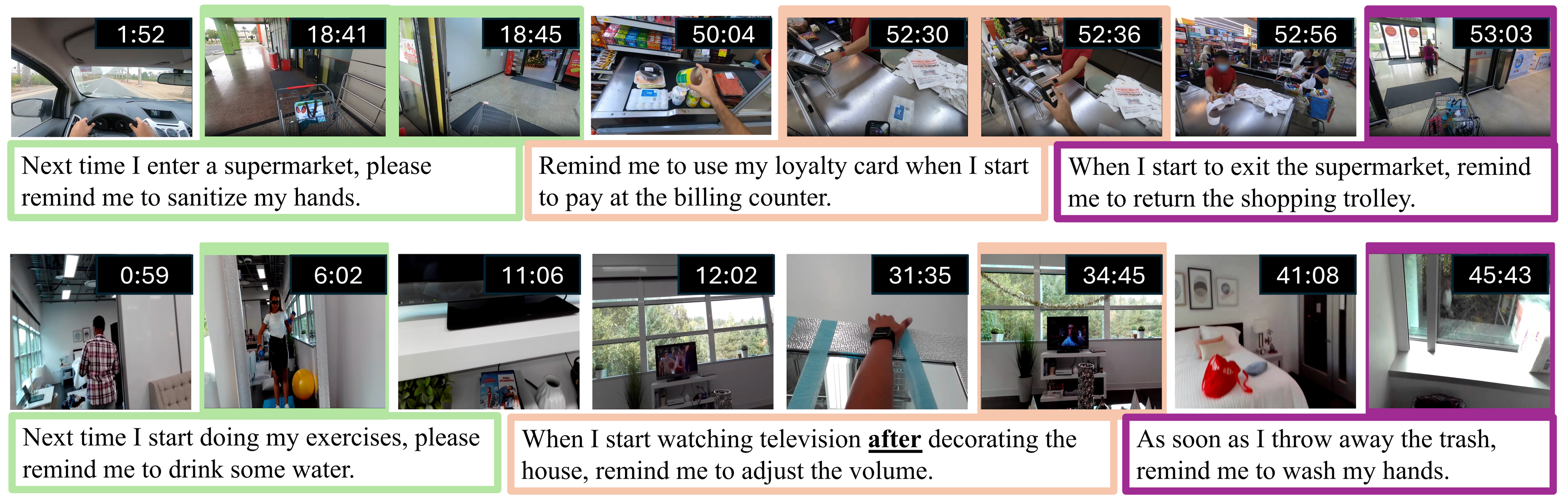
Project Website sdqesdataset.github.io

Motivation

Real-time applications like robotics and augmented reality need to swiftly detect and respond to **complex events** as they unfold, beyond a limited set of predefined classes.



We introduce a novel task called **Streaming Detection of Queried Event Start (SDQES)**, which leverages natural language to enable complex events descriptions in **streaming video**.



The goal of goal of SDQES is to output a high accuracy prediction of event start (*i.e.*, output time with **high accuracy and low latency**. We propose new metrics especially suited for measuring progress on this task: Streaming Recall and Streaming Minimum Distance.

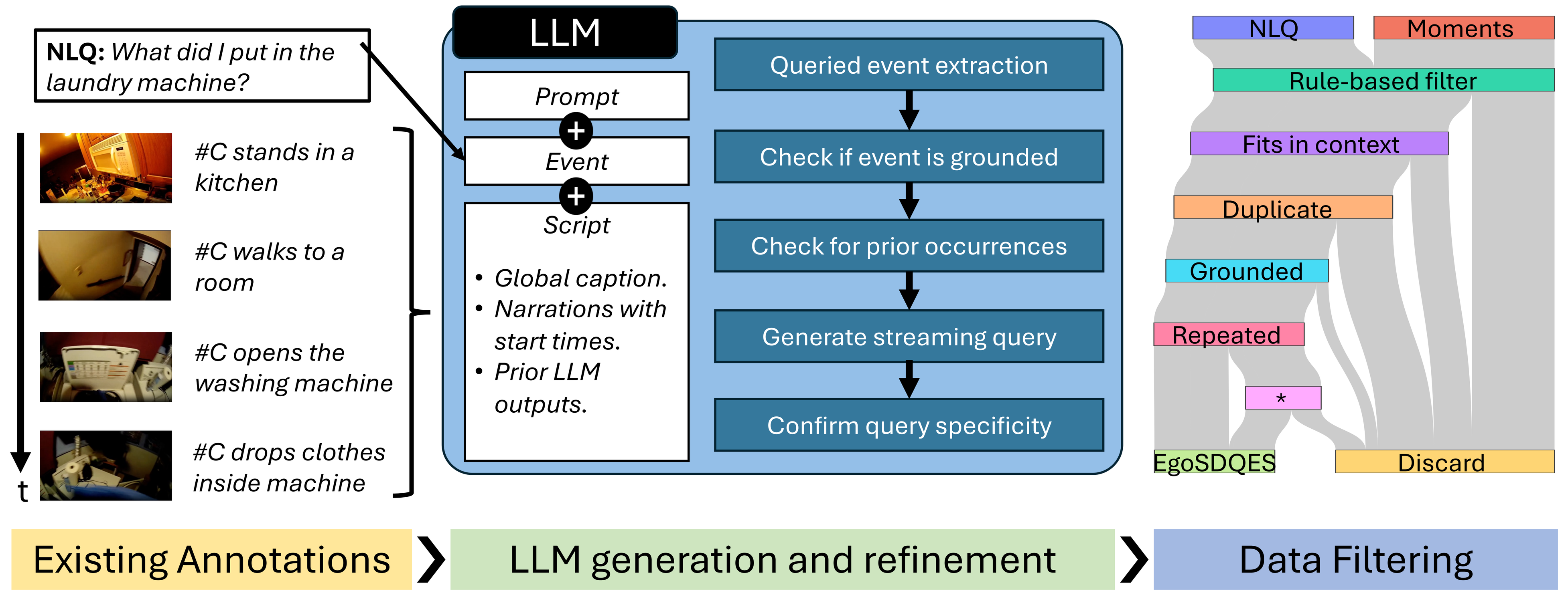
$$\mathrm{SR}(k, W) = \frac{1}{|Q|} \sum_{q \in Q} \mathbf{1}\{\exists t_{\text{out}} \in P_M^{(k)} : -anticipation \leq t_s - t_{\text{out}} \leq latency\}$$

$$\mathrm{SMD}(k) = \frac{1}{|Q|} \sum_{q \in Q} \min_{t_{\text{out}} \in P_M^{(k)}} |t_s - t_{\text{out}}|$$

The Challenge: there are **no existing datasets** for this task, and current ODAS **streaming models are limited**—they can only effectively output to a limited set of predefined classes.

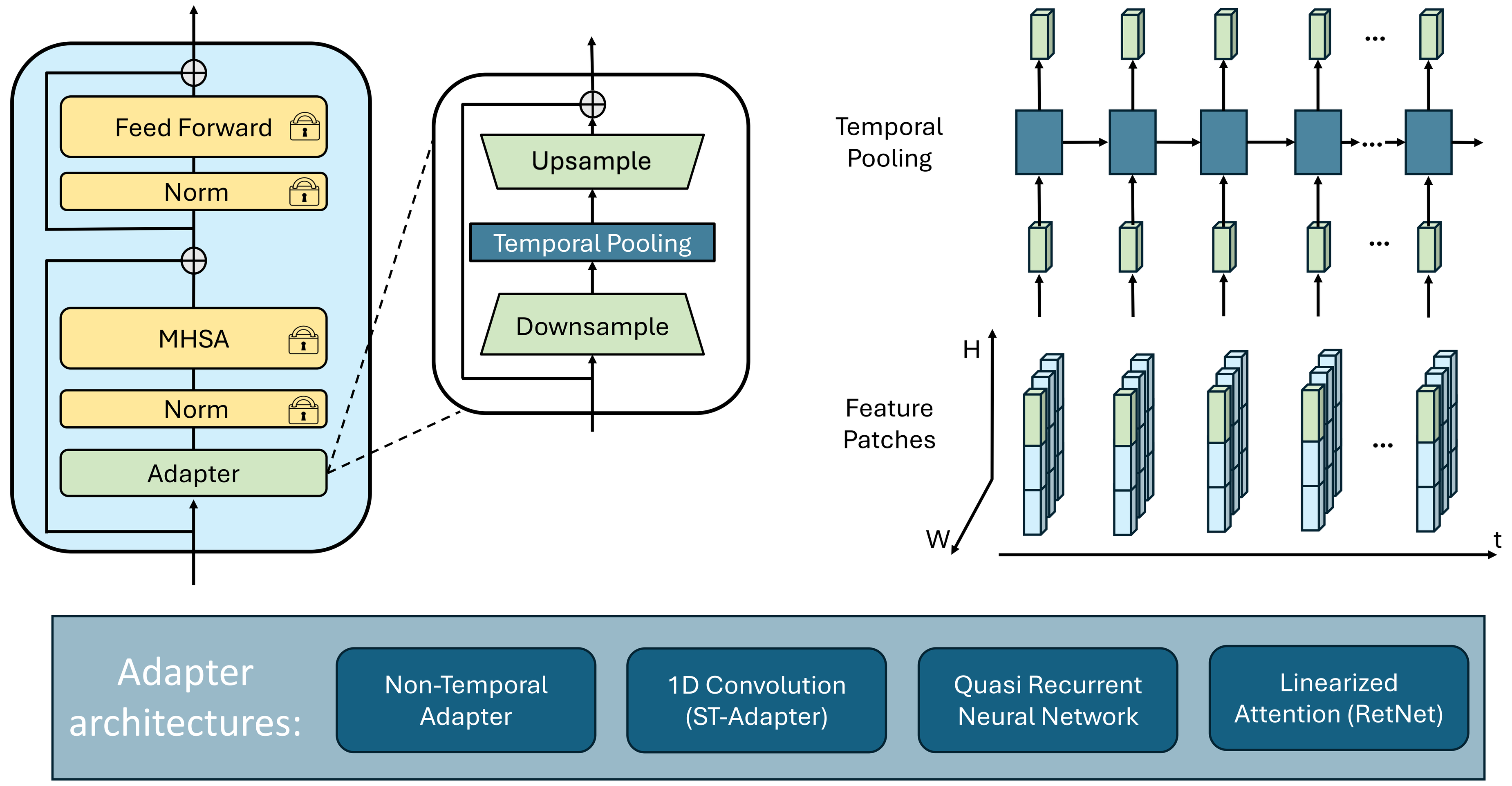
Dataset Collection

To address this, we develop a pipeline to generate a **new dataset for the task**, leveraging the Ego4D annotations to facilitate training and benchmarking of models for this capability.



Models: Streaming Adapters

Our proposed approach leverages pretrained vision-language foundation models by integrating **parameter-efficient streaming adapters** to deliver real-time, continuous event detection on **untrimmed video streams**.



We evaluate a **variety of combinations of Streaming Adapters and dual-encoder vision-language models**, including the current state-of-the-art (SOTA) egocentric video encoder.

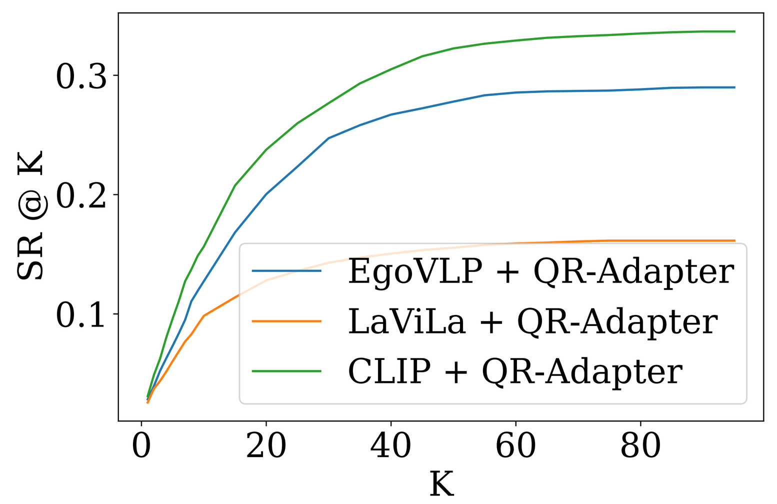
Main Results

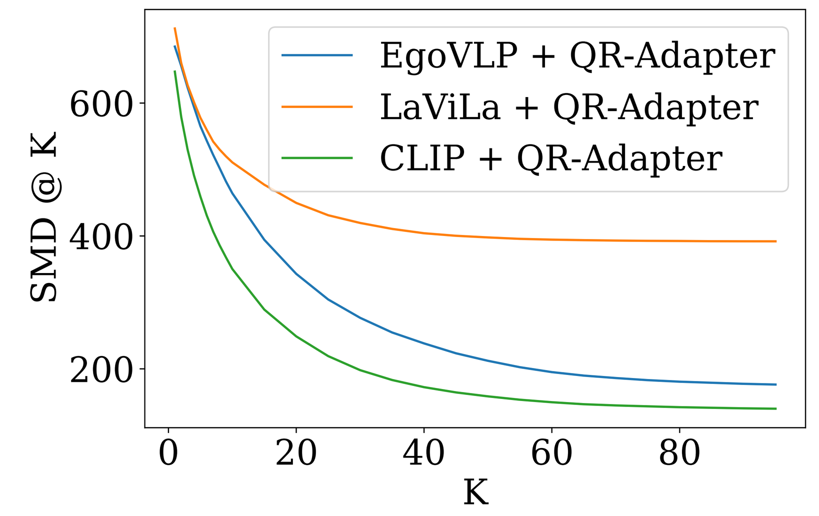
We **evaluate baselines** on both short clips and extremely long untrimmed videos.

Main takeaways include:

- **Training on our dataset significantly improves performance** on SDQES, with all adapted models finetunned on the generated data surpassing the zero-shot baseline.
- Streaming adapters with long temporal horizons outperform non-temporal models, proving that more **complex temporal modeling capabilities are beneficial** for SDQES.

	1 Min.		5 Min.					
Method	SR@1↑	SMD@1↓	SR@1↑	SR@2↑	SR@3↑	SMD@1↓	SMD@2↓	SMD@3↓
Zero-Shot CLIP	16.9	24.3	7.9	11.6	14.0	151.3	140.3	132.6
CLIP + Adapter	19.5	23.5	8.9	13.7	17.2	135.7	121.7	113.3
CLIP + QR-Adapter	23.7	21.2	9.1	14.1	18.7	136.7	117.7	102.9
LaViLa + Adapter	19.5	23.4	8.7	13.0	16.2	163.4	151.7	144.0
LaViLa + QR-Adapter	29.1	18.1	9.3	12.8	16.5	132.1	115.9	104.1
EgoVLP + Adapter	18.1	24.0	8.4	13.0	16.7	160.8	148.7	141.5
EgoVLP + QR-Adapter	28.8	17.7	9.7	14.1	17.9	133.1	120.8	110.9
EgoVLP + ST-Adapter	17.4	30.5	8.6	13.4	17.0	170.7	161.4	155.6
EgoVLP + RN-Adapter	25.7	21.3	9.4	15.4	20.1	174.8	159.0	149.2
EgoVideo + Adapter	27.1	28.8	16.0	21.8	26.4	148.5	138.3	131.2





- The **proposed models are efficient**, maintaining high performance with minimal latency, suitable for real-time applications.

	Memory	*Computational Latency*		
Model	# parameters	Multiply Adds	Floating Point Operations	Latency
EgoVLP backbone	180.92 M	7.85 TMACs	15.7 Tflops	1.68 s
EgoVLP + Adapter	+7.9%	+12.7%	+12.8%	+15.5%
EgoVLP + ST Adapter	+7.9%	+12.7%	+12.8%	+18.5%
EgoVLP + QRNN Adapter	+7.5%	+12.0%	+12.2%	+21.5%
EgoVLP + RetNet Adapter	+7.6%	+15.2%	+15.3%	+99.5%
EgoVLP Sliding Window	+0.1%	+298.5%	+298.8%	+260.2%