Self-Guided Consistent Sample Mining

Training Set D

Q: *Does <event> happen from <timestamps> in the video?*

Ego *Exo*

Video-LLM

Update

Discard inconsistent prediction samples

Transfer consistent prediction samples!

Predictions

D_{cons}

Dual-Encoder architecture

Target Answer $\mathbf{X_a}$

: Frozen *: Learnable*

Video-LLM

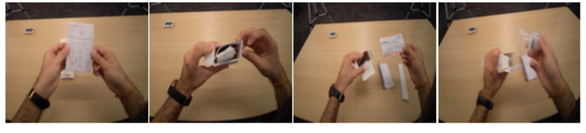
LoRA

$\mathbf{X_{ego}}$

Projection f_{ego}

Visual Encoder

Egocentric Video

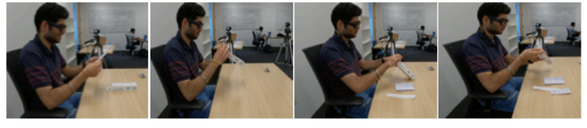


$\mathbf{X_{exo}}$

Projection f_{exo}

Visual Encoder

Exocentric Video



$\mathbf{X_q}$

Agent-centric Query

A person unboxes and arranges the covid test pack on the table.

Object-centric Query

The covid test pack is unboxed and arranged on the table.

Query

Unbox and arrange the covid test pack on the table