Modeling Fine-Grained Hand-Object Dynamics for Egocentric Video Representation Learning

Baoqi Pei[1,2], Yifei Huang[3,2], Jilan Xu[4,2] ,Guo Chen[4,2], Yuping He[5], Lijin Yang[3],
Yali Wang[6,2], Weidi Xie[7,2], Yu Qiao[2], Fei Wu[1], Limin Wang[5,2]

[1] Zhejiang University [2] Shanghai Artificial Intelligence Laboratory [3] The University of Tokyo
[4] Fudan University [5] Nanjing University[6] SIAT [7] Shanghai Jiao Tong University

ICLR

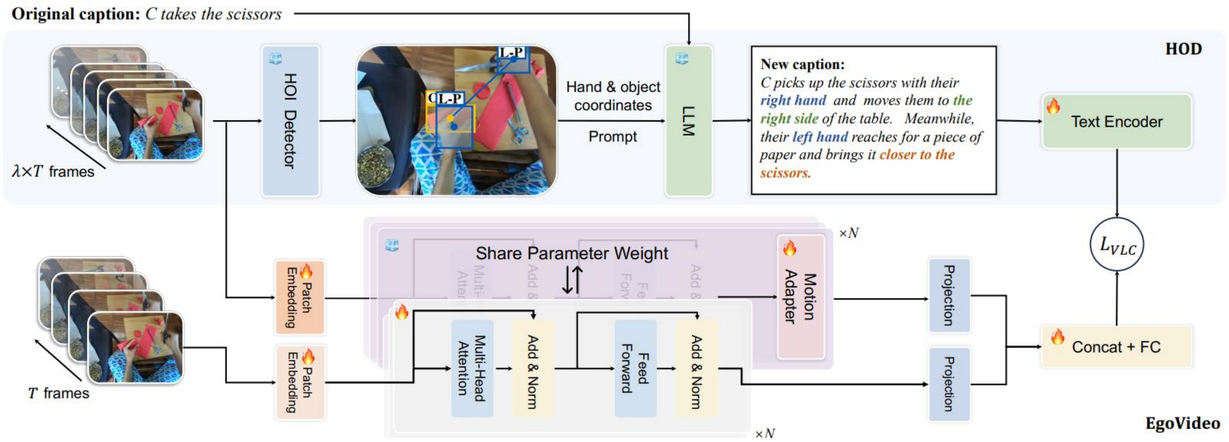
I. Introduction

Motivation: Original annotations in egocentric video datasets are highly condensed, neglecting a crucial aspect —— the fine-grained dynamics of hands and objects.

Contribution: We propose a new framework to integrate the modeling of fine-grained hand-object dynamics into the video representation learning process:

- **HOD data pipeline** to generate captions that describe fine-grained hand-object dynamics.
- **EgoVideo model** with a novel lightweight motion adapter and a co-training strategy.
- Get SOTA performance on **12** downstream tasks and robot manipulation tasks.

II. Framework



Data pipeline HOD:

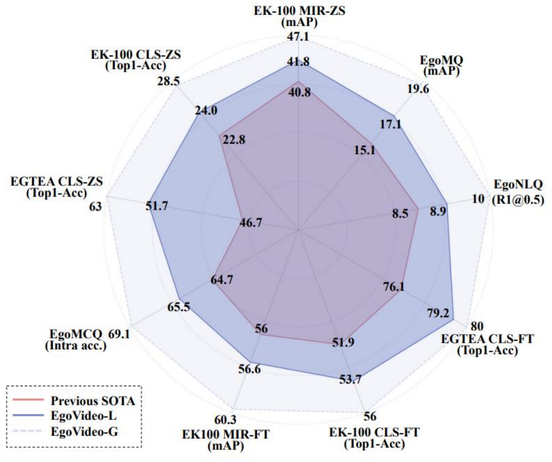
- Use a hand object detector to generate bounding boxes for hands and objects.
- Modified the boxes to a format that LLM understands.
- Use a LLM as Rephraser to enrich the original video captions.

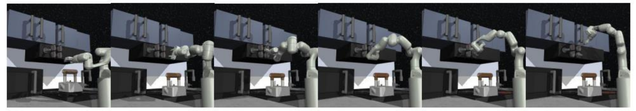
Egocentric Representation Learning Model: Egovideo

- Comprising a backbone and a motion adapter.
- Adopt a co-training strategy to obtain richer video representations

III. Experiments

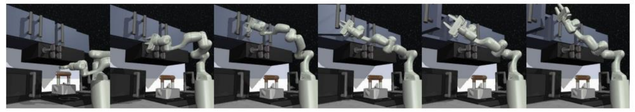
- Zero-shot and fine-tune results on multiple egocentric benchmarks.





Turn on the knob

Open microwave



Open door

- Qualitative results on the Franka Kitchen dataset.

We got 7 champions in the First Joint Egocentric Vision (EgoVis) Workshop!

Github Link:

Data & Training & Evaluation