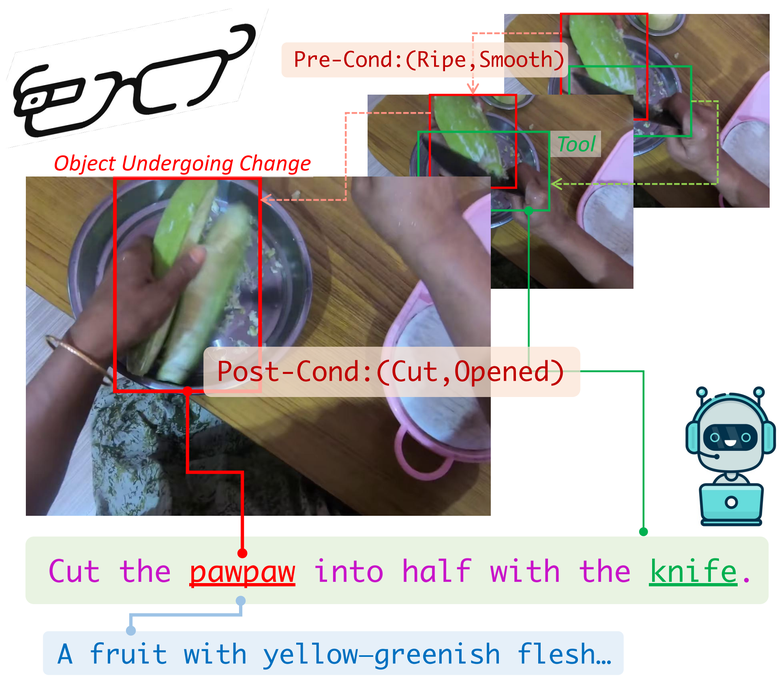
Localizing Active Objects from Egocentric Vision with Symbolic World Knowledge

Te-Lin (Albert) Wu, Yu (Bryan) Zhou*, Nanyun (Violet) Peng*
*University of California, Los Angeles (UCLA), *denotes equal contribution*

Introduction

What is "active object grounding", and why do we care about it?



OUC: Object undergoing **(state) change** due to an **action**
Tool: Object **facilitates** the action and causing the OUC state change

Pre-Condition: Pre-requisites to be met to enable the action to take place.
Post-Condition: Effects caused to and shown by the object being acted on.
Description: Describes the general attributes/appearances of the object.

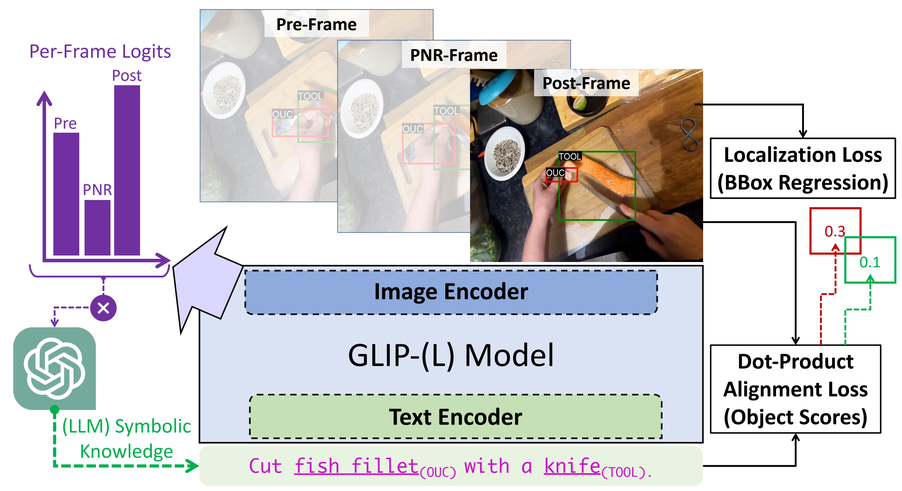
Our Main Contributions:

- *Task:* enriches the active object grounding task for next-gen. assistive AI.
- *Prompting:* devises a novel and useful LLM knowledge extraction pipeline.
- *Technique:* proposes an LLM-enhanced grounding techniques that emphasize on the symbolic world knowledge of action-object relations.

Method & LLM-Symbolic-World-Knowledge

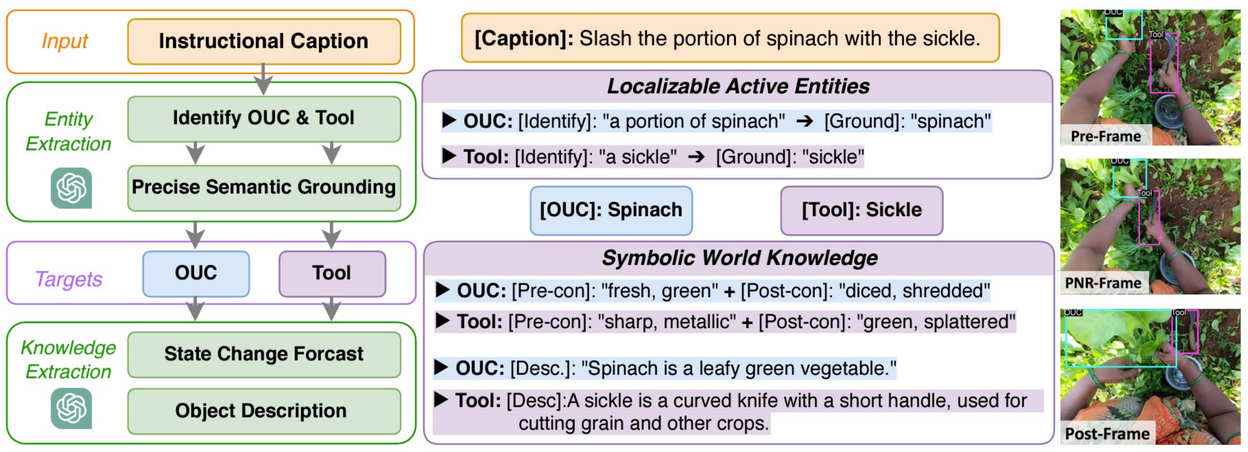
Model Overview:

1. Base architecture: GLIP model (an open vocabulary object detector).
2. LLM then enriches the ground-able knowledge.
3. Joint inference on weighting the temporal-dependent knowledge.

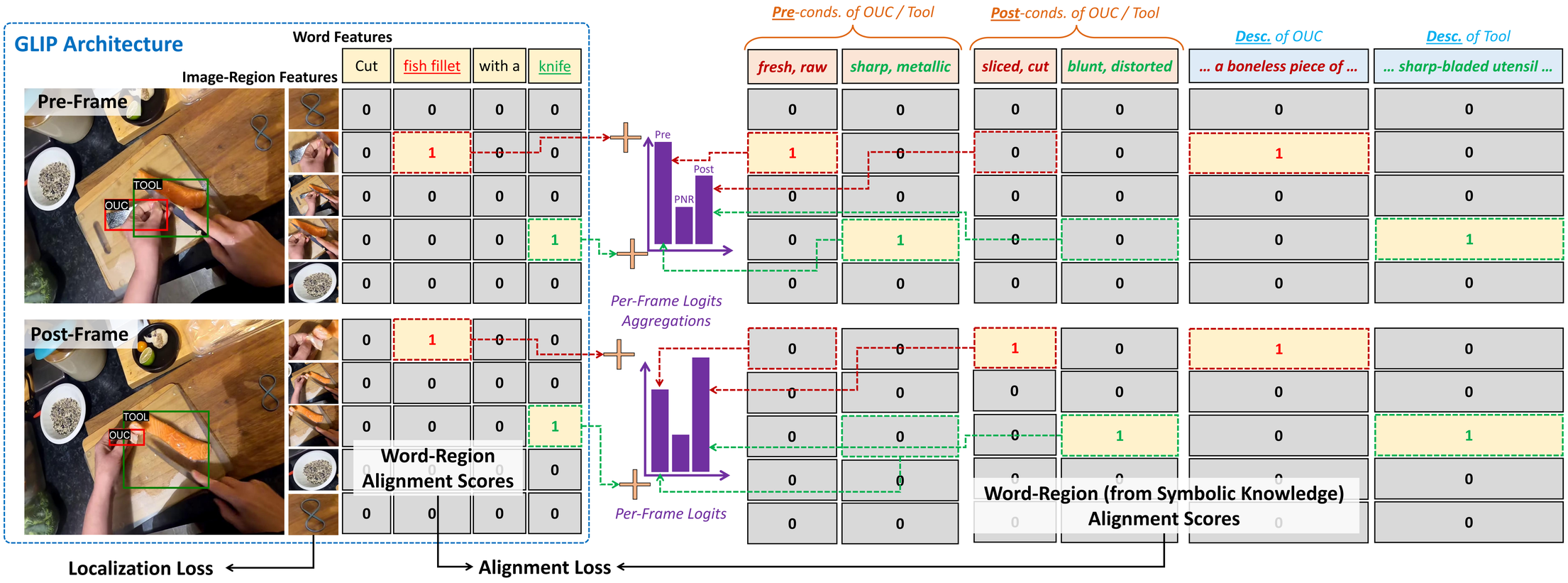


LLM Symbolic World Knowledge Extraction:

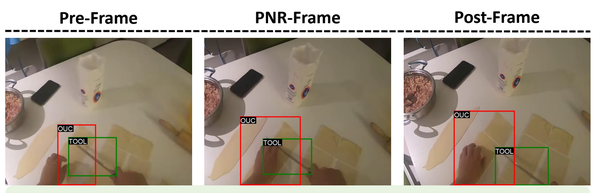
A novel prompting technique that is generalizable to many other relevant tasks.



Learning the LLM-enhanced Grounding



Task & Experimental Results (Quantitative)

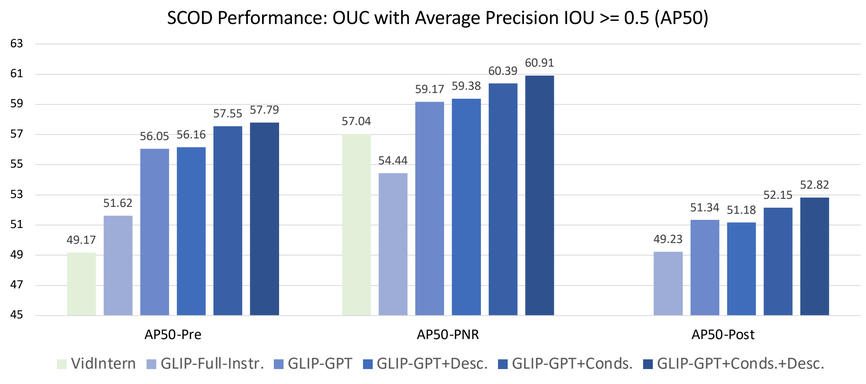


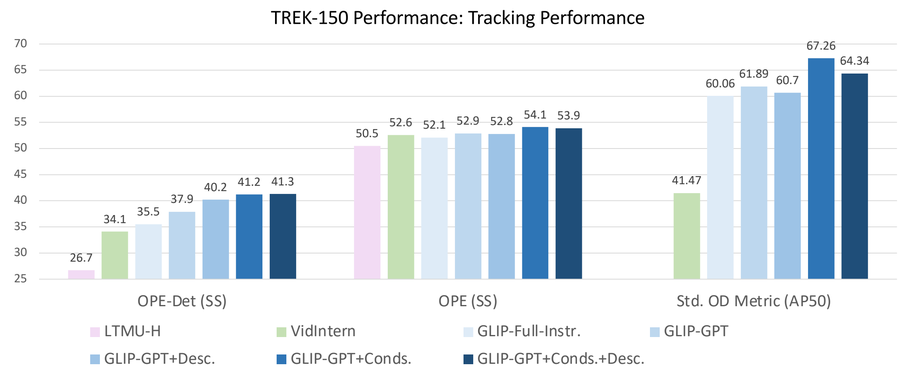
Tasks & Datasets:

1. Ego4D SCOD
2. Epic-Kitchens TREK-150

Metrics:

1. Average Precisions (AP50, AP75, AP[50:95:5])
2. One-Pass Evaluation (OPE) Success Score (SS)
   - OPE-Det: Use detection results for OPE





Qualitative Samples