SEIS 764: Artificial Intelligence
University of St. Thomas

Name: Khongmeng Kormoua

St. Thomas ID: 101323019

Literature Review

Title: Zero-shot HOI Detection with MLLM-based Detector-agnostic Interaction Recognition by Shiyu Xuan, Dongkai Wang, Zechao Li, Jinhui Tang.
https://arxiv.org/abs/2602.15124

Summary: "The paper proposes a zero-shot human–object interaction (HOI) detection method that leverages a Multimodal Large Language Model (MLLM) to recognize interactions without requiring training examples for every possible action. Instead of learning a fixed classifier for each interaction category, the approach separates object detection from interaction reasoning: any standard detector is used to locate humans and objects, and then an MLLM analyzes each human–object pair using visual features and language understanding to infer the interaction. Because MLLMs contain broad semantic and world knowledge, they can generalize unseen interactions by reasoning about plausible relationships between humans and objects. This detector-agnostic and reasoning-based framework significantly improves zero-shot HOI performance compared to traditional classification-based methods." Generated by ChatGPT on 2/20/2026

I personally am always interested in zero-shot and co-operation between multiple models architecture, as it suggests a team-play style of solving a problem and I believe it has high potential. Although, after reading into the paper, it is not actually what I expected, as I would think that MLLM could be used for object detection purposes too as it is capable of. But implementing in a way of modular, having plug-and-play capability and cost-efficiency is what the paper was about, as Zero-shot HOI detection has been covered before.

The paper mainly introduces a modular system and plug-and-play style architecture which I still believe has such high potential, as for my previous Literature Review also has similar concept. Because this can theoretically be applied to any industry. And the idea is very similar to how humans think, we think of an object which is presented on the scene and the interaction between them. This can be helpful in CCTV where an interaction is important such as a person is shopping or stealing, could AI tell?