ANNEXE: Unified Analyzing, Answering, and Pixel Grounding for Egocentric Interaction

https://yuggiehk.github.io/annexe/

Yuejiao Su, Yi Wang*, Qiongyang Hu, Chuang Yang, Lap-Pui Chau*

Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong SAR

JC STEM Lab of Machine Learning and Computer Vision 賽馬會「機器學習與計算機視覺」創科實驗室

Motivation

- Egocentric hand-object interaction can gain deeper insights into human interaction.
- Few existing studies have successfully integrated coherent **text-level** and fine-grained **pixel-level** responses as outputs.
- The responses of existing works are in fixed mode, which lacks **flexibility** when egocentric interaction results are employed for diverse downstream tasks.

RIS

Query: The red car.

EgoVQA

Query: What am I doing?

Answer: Holding a pink bucket.

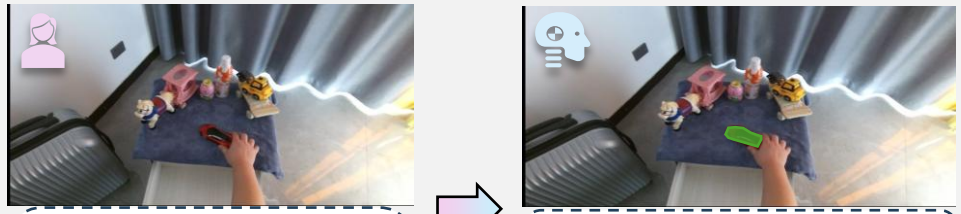
EHOI

< hand, hold, cup >

Contribution

- We present the new **Ego-IRG** task to interpret egocentric interaction comprehensively by a synergy of three ego-tasks: analyzing, answering, and pixel grounding.
- We propose a large-scale annotated **Ego-IRGBench** dataset.
- We present the **ANNEXE** for tackling the Ego-IRG task utilizing MLLMs.

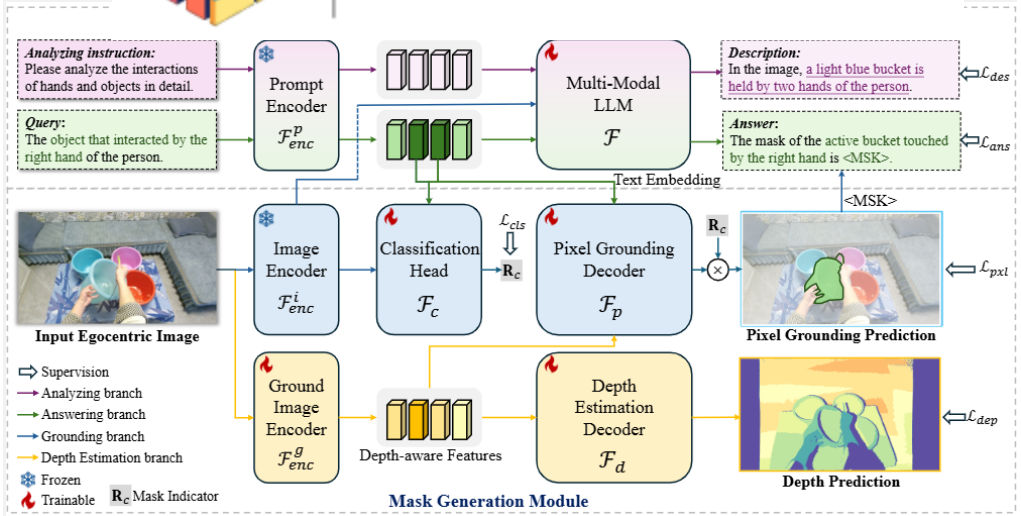
Proposed Ego-IRG Task



Query: Pleas in th activ

Analysis: A toy car is held by the right hand of the person in this

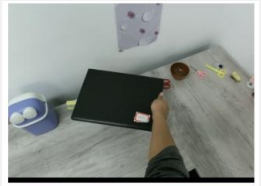
JC STEM Lab of Machine Learning and Computer Vision 賽馬會「機器學習與計算機視覺」創科實驗室



The first model that can under-stand visual-language inputs and generate text- and pixel-level responses regarding egocentric interactions.

Large-scale Ego-IRGBench Dataset

This dataset contains interaction descriptions for over 20k egocentric images and1.6M query-answer-mask paired labels.



Egocentric RGB Image

Egocentric Depth Map

Query 1: What is held by hand?

Answer 1: The toy car is held by hand, and the mask of it is <MSK>.

Example 1

Query n: The hands involving interactions.

Answer n: The toy car is held by the right hand, and the mask of it is <MSK>.

Example n

Experimental Results

Image + Query | Prediction | GT