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

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

{yuejiao.su,qiongyang.hu}@connect.polyu.hk,{yi-eie.wang,c1yang,lap-pui.chau}@polyu.edu.hk

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

## Abstract

*Egocentric interaction perception is one of the essential branches in investigating human-environment interaction, which lays the basis for developing next-generation intelligent systems. However, existing egocentric interaction understanding methods cannot yield coherent textual and pixel-level responses simultaneously according to user queries, which lacks flexibility for varying downstream application requirements. To comprehend egocentric interactions exhaustively, this paper presents a novel task named Egocentric Interaction Reasoning and pixel Grounding (Ego-IRG). Taking an egocentric image with the query as input, Ego-IRG is the first task that aims to resolve the interactions through three crucial steps: analyzing, answering, and pixel grounding, which results in fluent textual and fine-grained pixel-level responses. Another challenge is that existing datasets cannot meet the conditions for the Ego-IRG task. To address this limitation, this paper creates the Ego-IRGBench dataset based on extensive manual efforts, which includes over 20k egocentric images with 1.6 million queries and corresponding multimodal responses about interactions. Moreover, we design a unified ANNEXE model to generate text- and pixel-level outputs utilizing multimodal large language models, which enables a comprehensive interpretation of egocentric interactions. The experiments on the Ego-IRGBench exhibit the effectiveness of our ANNEXE model compared with other works.*

## 1. Introduction

Understanding how humans interact with the environment or objects through visual information has always been a pivotal focus in computer vision, which can contribute to constructing next-generation agent systems. Traditionally, studies in this area have concentrated on exocentric or third-person view (TPV) interaction analysis, achieving substan-
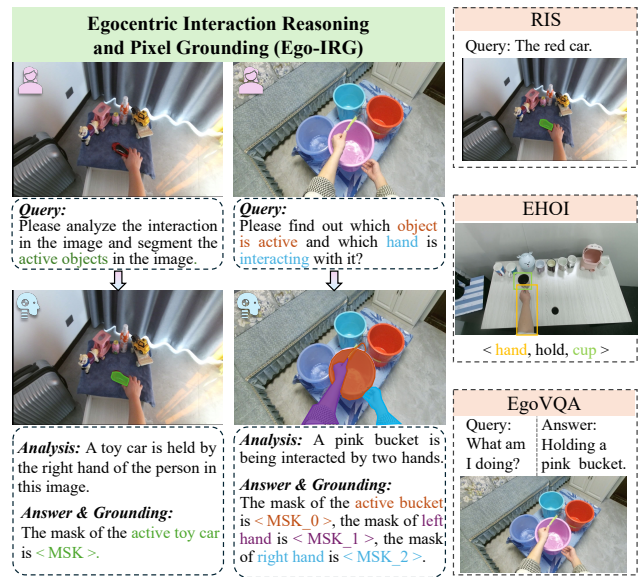


Figure 1. Ego-IRG task versus other related tasks. The Ego-IRG task advances the interaction understanding compared with other related tasks such as Referring Image Segmentation (RIS), Egocentric Hand-Object Interaction detection (EHOI), and Egocentric Question Answering (EgoVQA).

tial progress through evolving various tasks such as human-object interaction detection (HOI-det) [30, 53] and action recognition [11, 82]. Recently, a growing number of users spontaneously post videos and images taken from egocentric or first-person view (FPV) [34] on social media platforms due to the development of head-mounted devices (HMD) and GoPro. These egocentric data provide a unique perspective by showing visual signals from the perspective of the camera wearer, emphasizing individual behavior in different situations. The study of hand-object interaction from an egocentric perspective can gain deeper insights into human interaction, which has a wide range of applications such as augmented reality (AR) [8], embodied AI [66], and industrial human-assistive systems [32].

---

*Corresponding author.

However, severe occlusions and background variations pose significant challenges to FPV interaction study and create a huge domain gap between TPV and FPV, making it non-trivial to directly transfer exocentric algorithms to egocentric interaction analysis [36]. With the emergence of various large-scale egocentric datasets such as Ego4D [19], HOI4D [48], and EPIC-KITCHENS [13, 14], some recent works utilized egocentric visual information to detect [33], segment [64, 83], and recognize the interactions [70] between hands and objects. These studies aim to generate bounding boxes or masks for hands and objects while classifying action categories. Parallel to this, other research efforts have focused on producing egocentric video narrations [62, 75], which provide the text responses to describe the sequential interaction in detail. Although these advancements have significantly contributed to understanding human-object interaction from an egocentric perspective, few studies have successfully integrated coherent text-level and fine-grained pixel-level responses as outputs. These multi-modal outputs are all essential for a thorough comprehension of egocentric interactions. Furthermore, the responses of existing works are in fixed mode, which lacks flexibility when egocentric interaction results are employed for diverse downstream tasks.

In this paper, we propose a novel task called **Ego**centric **I**nteraction **R**easoning and pixel-**G**rounding (**Ego-IRG**) to facilitate the study of comprehensive egocentric interaction. As shown in Fig. 1, this task allows for unified analyzing, answering, and pixel grounding regarding interactions within egocentric images based on various user queries. Specifically, given a text query and an egocentric image, the Ego-IRG task encompasses two essential steps for comprehensive interaction parsing: reasoning and pixel grounding. The reasoning step aims to generate the text-level responses for interaction interpretation, which includes two sub-tasks: analyzing and answering. The analyzing sub-task aims to generate an overall text description of interactions depicted in the image, while the answering sub-task is to reply to user queries that the specific objects need to be segmented. Furthermore, the pixel grounding step aims to locate objects relevant to the query using fine-grained masks. The challenging Ego-IRG task unifies several egocentric vision tasks that are commonly addressed separately, such as referring image segmentation (RIS) [44, 74], egocentric hand-object interaction detection (EHOI) [33], and egocentric visual question answer (EgoVQA) [15]. By exploring this novel Ego-IRG task, egocentric interactions can be studied more systematically and query-oriented, resulting in flexibility when applied to real-world applications.

Nevertheless, the existing datasets cannot meet the requirements for the Ego-IRG task. To address this issue and facilitate subsequent studies, this paper constructs a large-scale dataset, **Ego**centric **I**nteraction **R**easoning and Grounding **Bench**mark (**Ego-IRGBench**). Built upon the HOI4D [48] dataset, we employed extensive manual efforts to relabel egocentric images with interaction descriptions and query-answer-mask pairs. The Ego-IRGBench comprises 20,681 egocentric images, each accompanied by descriptions depicting the interactions. In addition, over 1.6 million query-answer pairs, along with pixel-level grounding results, are included in the Ego-IRGBench.

Moreover, few studies have utilized depth information to enhance understanding of egocentric interactions. We argue that depth information is essential and helpful to distinguish the active objects. Specifically, when humans perform an action, they typically position the object of interest in the foreground, creating a notable depth distinction between hands and manipulating objects (foreground) relative to the background. The depth differences can be utilized to recognize the interacting objects, especially in scenarios where multiple objects of similar appearance and categories exist. Therefore, this paper introduces a unified depth-assisted **AN**alyzing, a**N**sw**E**ring, and pi**X**el grounding for **E**gocentric interaction (**ANNEXE**) network to tackle the novel Ego-IRG task. This network incorporates the text generation module and the mask generation module to tackle the reasoning and pixel grounding simultaneously. In addition, due to the success of multimodal large language models (MLLMs), we utilize MLLMs in the reasoning module to enhance the representations of queries and images. To summarize, our work has three main contributions:

- To facilitate the study of human behavior, we present the new Ego-IRG task to interpret egocentric interaction comprehensively by a synergy of three ego-tasks: analyzing, answering, and pixel grounding.
- To address benchmark limitations, we propose a large-scale annotated Ego-IRGBench dataset containing interaction descriptions for over 20k egocentric images and 1.6M query-answer-mask paired labels.
- We present the ANNEXE for tackling the Ego-IRG task utilizing MLLMs, which is the first model that can understand visual-language inputs and generate text- and pixel-level responses regarding egocentric interactions.

## 2. Related Work

This paper mainly focuses on interaction understanding from an egocentric perspective and leverages the MLLMs to establish the ANNEXE model. Therefore, we elaborate the related work from three aspects: egocentric interaction understanding, vision-language foundation models, and multi-model egocentric understanding.

**Egocentric interaction understanding.** Some existing works perform action recognition [36, 70] to understand egocentric interactions coarsely, i.e., classify verbs and nouns of interactions in egocentric videos or images.

To improve recognition performance, action-relevant cues [27, 29, 41, 60, 71, 72, 84] such as hands, active objects, and their relations are considered to concentrate on executing interactions. For example, Kown et al. [27] first proposed the H2O dataset and benchmark to predict hand pose and object pose for egocentric interaction recognition. Wang et al. [60, 71, 72] utilized active object detection to concentrate more precisely on the occurring interaction. Furthermore, Shiota et al. [35, 47, 50, 63, 71, 79] considered the hand-object contact and object status to recognize interaction accurately. Parallel to these, some other researchers utilized different modalities [46, 61, 65, 80], such as gaze [22] and audio [18], to improve the recognition performance further. For example, Wang *et al.* [73] leveraged the actor motion to improve the performance of classifying verbs and nouns.

Other recent existing works focus on detecting or segmenting the hands and interacting objects in an egocentric view. For example, Bambach et al. [3] introduced a simple candidate region generation approach to locate and distinguish different hands and activities. Zhang et al. [64, 83] proposed to segment the active objects with the contacting hands. Leonardi et al. [31] proposed a method that detects hands and objects in the scene and determines which objects are currently involved in an interaction. **Remark**: These methods have completed the coarse understanding of egocentric interactions, but they lack a comprehensive understanding of interaction, like fluent text and fine-grained mask responses. In addition, the results cannot be directly applied to multiple downstream tasks with different requirements or queries.

**Vision-language foundation models.** Recently, large language models (LLMs) [1, 7, 21, 40, 54, 55, 69] have successfully achieved excellent performance on multiple language tasks due to their strong capabilities in understanding context. Consequently, vision-language models (VLMs) [2, 37, 45, 85] have emerged by leveraging the advanced capabilities of large language models (LLMs). By combining linguistic and visual processing skills, VLMs can perform a range of tasks, such as visual question answering [38, 52, 56], visual reasoning [39, 49, 57], and image retrieval [78]. These models are capable of generating natural language descriptions based on input images, responding to image-related questions, and retrieving relevant images based on textual queries.

MLLMs have also been widely used in visual grounding, which aims to identify objects in images according to given language descriptions. For example, Rasheed et al. [59] proposed GLaMM to produce natural language responses intricately linked with corresponding object segmentation masks. By combining the language generation capabilities of MLLMs with segmentation, Lai et al. [28] demonstrated a large language-instructed segmentation assistant, LISA, that can manage complex reasoning tasks. In addi-

tion, Yang et al. [77] created the LAVT to integrate linguistic and visual features in the middle layers of a vision Transformer encoder, enhancing cross-modal alignment and producing precise segmentation outcomes with a lightweight mask predictor.

**Multi-model egocentric understanding.** Previous works have also attempted to combine visual and language information using MLLMs from an egocentric perspective. Specifically, some researchers concentrated on egocentric video-language pretraining. For example, Lin et al. [43] created the first video-text pretraining EgoClip dataset containing the clip-text pairs. Based on this dataset, Valdez et al. [67] presented SViTT-Ego to solve the memory bottleneck problem existing in pertaining.

In addition, some people comprehended the content in egocentric images and videos through understanding and answering textual questions. For example, Jia et al. [6, 10, 15, 23] carefully crafted questions to comprehend human actions and intents and proposed egoTaskQA to support the model in responding to these questions in egocentric videos. Chen et al. [10] proposed the multi-hop video question-answering task to provide temporal supporting evidence for further reasoning. A small number of studies also combined egocentric action or interaction with VLM. For example, Kim et al. [25] proposed the PALM, which uses VLM to generate narrations and action sequences of past videos and predict future action sequences based on the generated content. In addition, Cheng et al. [12] proposed EgoThink to infer six tasks, including activity and localization. The most relevant work with this paper is HOI-ref [5], which uses language prompts to infer relevant content using bounding boxes regarding interaction. However, this work ignores analyzing the interaction and fails to achieve natural and fluent language output and pixel-level mask responses.

**Remark**. Although these methods have achieved coarsely visual grounding (bounding boxed) and/or text responses (verbs, nouns) for egocentric interactions, two problems still remain. First, they cannot generate specific outputs according to different queries, which limits the applications of egocentric interaction understanding in a variety of downstream tasks with distinct requirements. Second, existing works fail to offer a thorough interpretation since they cannot incorporate text- and pixel-level responses to understand egocentric interactions.

## 3. Methodology

This paper presents a novel task called Ego-IRG, which aims to generate both the text- and pixel-level responses to the user query about the interaction depicted in the egocentric image. The Ego-IRG task consists of three progressive sub-tasks, i.e., analyzing, answering, and pixel grounding. We begin this section by formally defining the Ego-IRG task in Sec. 3.1. In Sec. 3.2, we introduce the proposed
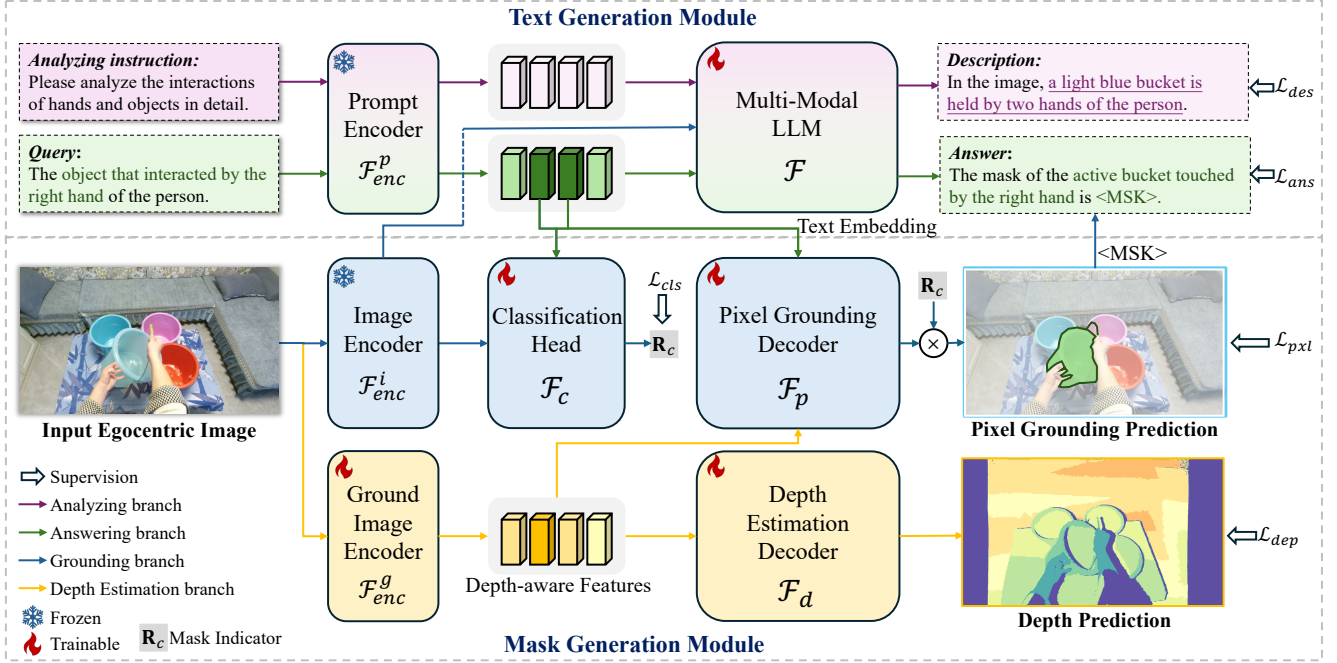
Figure 2. Overall architecture of the proposed ANNEXE model, a synergy of text generation and mask generation modules.

model ANNEXE in detail. Finally, the illustration of the proposed Ego-IRGBench is demonstrated in Sec. 3.3.

### 3.1. Task Definition

To facilitate the study of human actions, this paper proposes a novel Ego-IRG task, which unifies analyzing, answering, and pixel grounding regarding interactions in egocentric images based on user queries. Taking an egocentric image $\mathbf{I}$ and a query $\mathbf{T}$ as input, the primary goal of the Ego-IRG task is to analyze the interaction in $\mathbf{I}$ and produce the interaction description $\mathbf{R}_D$. Subsequently, the Ego-IRG task involves generating the answers $\mathbf{R}_A$ containing the objects that need to be segmented in response to the query, along with the pixel-level mask grounding results $\mathbf{R}_M$ of the object participating in the interaction specified in the query.

### 3.2. ANNEXE Architecture

To address the proposed Ego-IRG task, this paper establishes a unified ANNEXE model, the first model that can generate text-level and fine-grained pixel-level responses simultaneously for egocentric interaction understanding. As depicted in Fig. 2, the ANNEXE model can be divided into two modules: text generation and mask generation. Specifically, taking an egocentric image $\mathbf{I}$ and a query $\mathbf{T}$ as input, the text generation module is used to generate the textual responses, including the interaction description $\mathbf{R}_D$ and answers $\mathbf{R}_A$ for the query. Moreover, the mask generation module is designed to predict the fine-grained pixel-level grounding results $\mathbf{R}_M$ according to the query, which shows the location and boundary for the referring objects. Effectively integrating two modules, the ANNEXE model can complete the analyzing, answering, and pixel-grounding sub-tasks for comprehensive egocentric interactions. These two modules will deliberated as follows.

#### 3.2.1. Text Generation Module

As illustrated in Fig. 2, the text generation module mainly focuses on tackling two sub-tasks of Ego-IRG: analyzing and answering, which requires generating the interaction description and answers of the query. Due to the significant success of the MLLMs, we utilize their strong capabilities to understand and align the context of language and image in the text generation module to predict text responses. A typical approach is to use the prompt and image encoders to extract the textual and visual features of queries and images for prediction. However, in our case, the analyzing sub-task lacks an explicit text prompt, which affects the confidence of generated interaction descriptions. Hence, we introduce a hidden analyzing instruction $\mathbf{T}_a$ to guide the MLLM to predict interaction description $\mathbf{R}_D$ precisely. Specifically, taking the query $\mathbf{T}$ and hidden instruction $\mathbf{T}_a$ as inputs, we employ a frozen prompt encoder [9] to extract their features. In addition, we use a frozen image encoder [16] to extract the representative features of the input egocentric image $\mathbf{I}$. The process of extracting the features can be represented as:

$$\mathbf{F}_{enc}^{img} = \mathcal{F}_{enc}^i(\mathbf{I}, W_{enc}^i), \tag{1}$$

$$\mathbf{F}_{enc}^{ins}, \mathbf{F}_{enc}^{que} = \mathcal{F}_{enc}^p(\mathbf{T}_a, \mathbf{T}, W_{enc}^p), \tag{2}$$

where the $\mathbf{F}_{enc}^{img}$, $\mathbf{F}_{enc}^{ins}$, and $\mathbf{F}_{enc}^{que}$ are the extracted features of the egocentric image, analyzing instruction, and input query, respectively. And the functions $\mathcal{F}_{enc}^{i}(\cdot)$ and $\mathcal{F}_{enc}^{p}(\cdot)$ represent the image encoder with the parameters of $W_{enc}^{i}$ and the prompt encoder with the parameters of $W_{enc}^{p}$ respectively.

After feature extraction, the trainable multi-modal LLM $\mathcal{F}$ is utilized to obtain two types of textual output, *i.e.*, interaction descriptions, and answers for the query about objects that need to be segmented. This process can be represented as:

$$\mathbf{R}_D, \mathbf{R}_A = \mathcal{F}(\mathbf{F}_{enc}^{img}, \mathbf{F}_{enc}^{ins}, \mathbf{F}_{enc}^{que}, W), \quad (3)$$

where the $\mathcal{F}(\cdot)$ means the multi-modal LLM with the parameter of $W$.

In practice, the MLLM we used is Mini-GPT4v2 [9], and the image and prompt encoders are frozen while the $\mathcal{F}$ is trainable. Through this text generation module, the MLLM can be trained to generate coherent and precise interaction descriptions and answers depending on different queries and instructions, which further enhances the capability of the model to understand the in-depth information of input texts and images.

### 3.2.2. Mask Generation Module

The mask generation module is incorporated to complete the pixel grounding sub-task, which requires obtaining fine-grained masks of queried objects in the egocentric image. To predict masks accurately, we observed that depth information is essential for pixel grounding about egocentric interactions. It is because when people perform an activity, the operating objects and hands are usually positioned in the foreground, which has a relatively clear distinction from the background in depth, especially when multiple similar objects are in the same scenario. Therefore, we regard the depth estimation as an auxiliary task to integrate the depth information into the network. As shown in Fig. 2, the depth estimation branch is also incorporated to help distinguish the foreground and background for pixel grounding in the mask generation module.

Specifically, taking the image feature $\mathbf{F}_{enc}^{img}$ and the query feature $\mathbf{F}_{enc}^{que}$ as the input, we first use a trainable classification head $\mathcal{F}_c$ to predict the mask indicator, which indicates whether the targeted objects are in the image. In this binary classification setting, if the predicted mask indicator is true, the pixel-level grounding mask will be generated further, while the mask prediction will not be considered if the mask indicator is false. In addition, we select some representative features $\mathbf{F}_{sel}^{que}$ from the query feature $\mathbf{F}_{enc}^{que}$ along with $\mathbf{F}_{enc}^{img}$ as inputs to the classification head $\mathcal{F}_c$. In practice, the classification head includes several stacked convolution-relu-pooling layers with multi-layer perceptron. The classification process is:

$$\mathbf{R}_c = \mathcal{F}_c(\mathbf{F}_{enc}^{img}, \mathbf{F}_{sel}^{que}, W_c), \quad (4)$$

where the $\mathbf{R}_c$ is the predicted mask indicator with the value of 0 or 1, which is the intermediate output of the mask generation module. The $\mathcal{F}_c(\cdot)$ means the classification head with the parameters of $W_c$.

Furthermore, we regard depth estimation as an auxiliary task that helps the network extract image features that integrate latent deep information. Specifically, we add an additional depth estimation branch in the mask generation module, in which a trainable ground image encoder $\mathcal{F}_{enc}^{g}$ is used to extract depth-aware image features $\mathbf{F}_{g}^{img}$. A trainable depth estimation decoder $\mathcal{F}_d$ with the parameters of $W_d$ is also utilized to predict the depth map of the input egocentric image, which encourages the ground image encoder to learn spatial information rather than appearance characteristics only. The process of depth estimation is as follows:

$$\mathbf{F}_{g}^{img} = \mathcal{F}_{enc}^{g}(\mathbf{I}, W_g), \quad (5)$$

$$\mathbf{R}_d = \mathcal{F}_d(\mathbf{F}_{g}^{img}, W_d), \quad (6)$$

where the $\mathbf{R}_d$ is the depth map prediction. And the $W_g$ is the parameter set of the ground image encoder $\mathcal{F}_{enc}^{g}$. In practice, we use the decoder in Depth Anything V2 [76] as the depth estimation decoder to predict the depth map.

Consequently, based on the selected query features $\mathbf{F}_{sel}^{que}$, the extracted ground image features $\mathbf{F}_{g}^{img}$, and the predicted mask indicator $\mathbf{R}_c$, the predicted mask will be generated through the trainable pixel grounding decoder $\mathcal{F}_p$. Specifically, taking $\mathbf{F}_{sel}^{que}$ and $\mathbf{F}_{g}^{img}$ as input, some convolutional and max-pooling layers are used to downsample these features into the same dimension. Then we add the $\mathbf{F}_{sel}^{que}$ and $\mathbf{F}_{g}^{img}$ with the same dimension to be the combined feature, which is sent into pixel grounding decoder $\mathcal{F}_p$ to predict the initial mask. Finally, the initial mask is multiplied with the predicted mask indicator $\mathbf{R}_c$ to obtain the final mask. This process can be denoted as follows:

$$\mathbf{R}_M = \mathcal{F}_p((f_{down}(\mathbf{F}_{g}^{img}) + f'_{down}(\mathbf{F}_{sel}^{que})), W_p) \times \mathbf{R}_c, \quad (7)$$

where the $\mathbf{R}_M$ is the generated mask, the $W_p$ is the trainable parameter of the pixel grounding decoder, the $\times$ means the multiplication of mask indicator and initial mask, and the $f_{down}(\cdot)$ and $f'_{down}(\cdot)$ are the down-sampling functions.

By incorporating the mask generation module, the proposed ANNEXE model can predict precise and query-orientated pixel-level masks regarding egocentric interactions, which can be applied to various downstream tasks with different requirements more efficiently.

### 3.2.3. Training

The training loss of the ANNEXE model consists of five parts, *i.e.*, the analyzing loss $\mathcal{L}_{des}$ for interaction description, the answering loss $\mathcal{L}_{ans}$ for the response to queries, the classification loss $\mathcal{L}_{cls}$ for mask indicator, the depth estimation loss $\mathcal{L}_{dep}$, and the pixel grounding loss $\mathcal{L}_{pxl}$ for
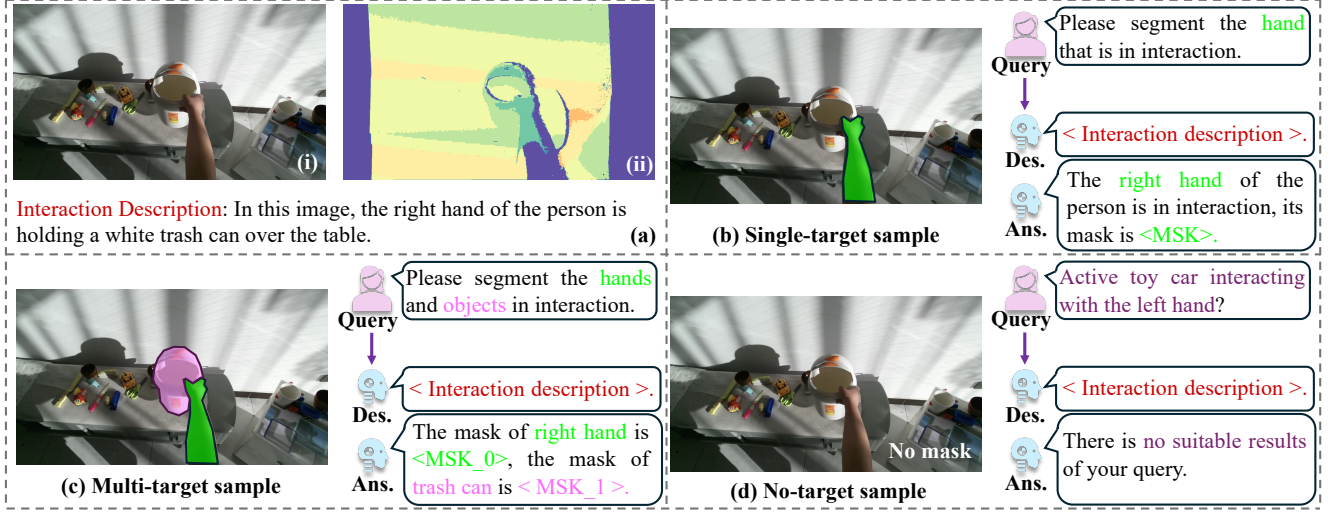
Figure 3. Illustration of the proposed Ego-IRGBench dataset, which includes (a)(i) the egocentric image, (a)(ii) the depth map, and interaction description. It also includes (b) single-, (c) multi-, and (d) no-target samples with the corresponding query, description (Des.), and answer (Ans.).

Table 1. Ego-IRGBench versus other datasets. Obviously, Ego-IRGBench is the first dataset that provides matching descriptions (Des.), answers (Ans.), and masks for the egocentric image-query inputs regarding interactions (Inter.).

| Dataset | Ego/ Exo | Inter. | Query | Responses | | |
|---|---|---|---|---|---|---|
| | | | | Des. | Ans. | Mask |
| SAM-1B [26] | Exo | ✗ | ✓ | ✗ | ✗ | ✓ |
| GranD [59] | Exo | ✗ | ✗ | ✓ | ✗ | ✓ |
| RefCOCO [24] | Exo | ✗ | ✓ | ✗ | ✗ | ✓ |
| RefCOCO+ [24] | Exo | ✗ | ✓ | ✗ | ✗ | ✓ |
| RefClef [24] | Exo | ✗ | ✓ | ✗ | ✗ | ✓ |
| G-Ref [51] | Exo | ✗ | ✓ | ✗ | ✗ | ✓ |
| VISOR [14] | Ego | ✓ | ✓ | ✗ | ✓ | ✓ |
| EPIC-KITCHENS [13] | Ego | ✓ | ✗ | ✓ | ✗ | ✗ |
| EGTEA [42] | Ego | ✓ | ✗ | ✓ | ✗ | ✓ |
| HOI4D [48] | Ego | ✓ | ✗ | ✗ | ✗ | ✓ |
| Ego4D [19] | Ego | ✓ | ✓ | ✗ | ✗ | ✗ |
| EgoHOS [83] | Ego | ✓ | ✗ | ✗ | ✗ | ✓ |
| VISOR-NVOS [62] | Ego | ✓ | ✗ | ✓ | ✗ | ✓ |
| **Ego-IRGBench** | Ego | ✓ | ✓ | ✓ | ✓ | ✓ |

final mask. We use the cross entropy loss for $\mathcal{L}_{des}$, $\mathcal{L}_{ans}$ [85], and $\mathcal{L}_{pxl}$. The binary cross entropy loss is used for $\mathcal{L}_{cls}$, and smooth L1 loss is utilized for $\mathcal{L}_{dep}$. Therefore, the overall loss function for the ANNEXE is:

$$\mathcal{L} = \lambda_{des}\mathcal{L}_{des} + \lambda_{ans}\mathcal{L}_{ans} \quad (8)$$
$$+ \lambda_{cls}\mathcal{L}_{cls} + \lambda_{dep}\mathcal{L}_{dep} + \lambda_{pxl}\mathcal{L}_{pxl}, \quad (9)$$

where the $\mathcal{L}$ is the final loss, and the $\lambda_{des}$, $\lambda_{ans}$, $\lambda_{cls}$, $\lambda_{dep}$, and $\lambda_{pxl}$ are the weight of the corresponding loss function.

### 3.3. Ego-IRGBench

We develop a new Ego-IRG task to provide a comprehensive framework for comprehending egocentric interactions, as shown in Sec. 3.1. The objective of the Ego-IRG task

is to generate the *interaction descriptions, answers, and pixel-level masks* simultaneously based on the input text queries and egocentric images. However, to the best of our knowledge, very few datasets can satisfy the requirements of Ego-IRG for the community to build up data-driven algorithms. Therefore, to encourage future studies of the Ego-IRG task, we present a high-quality Ego-IRGBench dataset. This dataset is built upon the HOI4D [48] dataset, and we use a stage-wise annotation pipeline that consumes extensive manual efforts to re-label the dataset with queries and matching text- and pixel-level responses. Unlike other related datasets, the Ego-IRGBench is the first to provide fluent textual responses and pixel-level grounding masks for understanding egocentric interaction. Tab. 1 displays the comparison of the Ego-IRGBench with other datasets.

**Structure and scale.** The Ego-IRGBench is a comprehensive dataset that includes over 1.6 million queries along with their corresponding text- and pixel-level response pairs. This extensive dataset is built upon a collection of 20,504 RGB-D egocentric image pairs extracted from the HOI4D dataset [48], including various interactions and environments from a first-person view. Specifically, each egocentric RGB image is paired with a depth map (as shown in Fig. 3 (a)(ii)) and an interaction description (Fig. 3), providing spatial information about the scene and outlining the specific interaction taking place. Furthermore, multiple queries are labeled for each image, allowing for various inquiries related to the interactions depicted. Each query is paired with an answer and corresponding pixel-level grounding mask, forming systematic and ultimate feedback to the query. It is worth emphasizing that the query is not limited to inferring a single target (Fig. 3 (b)) involving interactions. Multi-target (Fig. 3 (c)) and

no-target (Fig. 3 (d)) queries are also provided in the Ego-IRGBench dataset, which shows the diversity of the dataset. More explanations for the Ego-IRGBench dataset can be found in Appendix A.1.

**Step-wise annotation pipeline.** To efficiently build up the dataset that meets the requirements of the Ego-IRG task, we implemented a semi-automated annotation pipeline from coarse to fine. Specifically, the annotation pipeline includes three steps: interaction classification, hand-object mask generation, and query-response generation. In the first step, the employed experts are asked to classify the aligned RGB-D images to determine whether interactions are taking place. As a result, the images are divided into two categories: with and w/o interactions. For samples that include interactions, the description of the interaction is generated semi-manually. In the second step, masks for the hands and interacting objects are generated semi-automatically based on the masks of the HOI4D [48] dataset. Finally, queries and textual responses are created manually, while the corresponding pixel-level grounding masks are generated automatically. More details about annotations are explained in Appendix A.2.

**Evaluation criteria.** To build up the benchmark for Ego-IRG, we split the Ego-IRGBench into train, validation, and test sets according using a ratio of 5:2:3, yielding a train set of 10,249 images with 806,982 queries, a validation set of 4,094 images with 322,208 query-response-mask pairs, and a test set of 6,161 images and 485,174 query-response-mask pairs. In validation and testing, three key aspects are evaluated: i) generated interaction description quality, ii) quality of generated answers for queries, and iii) pixel-level grounding accuracy. To examine the qualities of generated interaction descriptions and answers, METEOR [4] and CIDEr [68] metrics are considered following previous work [59]. To verify the accuracy of pixel grounding, cIoU is considered as the metric following the work in [28, 74]. More details are illustrated in Appendix A.3.

# 4. Experiments

This paper proposes a unified Ego-IRG task for egocentric interaction understanding, along with the Ego-IRGBench dataset and ANNEXE model. We first provide the dataset that we used for the experiments in this section, followed by implementation details in Sec. 4.1. The effectiveness of the suggested ANNEXE is then demonstrated by the quantitative results, which are shown in Sec. 4.2. Finally, the qualitative results are displayed in Sec. 4.3 to visualize the results of ANNEXE for the Ego-IRG task.

## 4.1. Dataset and Implementation Details

As explained in Sec. 3.3, we proposed a comprehensive Ego-IRGBench dataset that consists of over 1.6 million query-response-mask pairs about egocentric interactions. This dataset is used for training and testing.

For network architecture, unless otherwise specified, we use Mini-GPT4v2 [9] as the base multi-modal LLM $\mathcal{F}$, and adopt the Depth Anything V2 [76] as the depth estimation branch. Additionally, we utilize the instruction template $[INST] \ <Img><ImageHere></Img>$ $[task] \ Instruction \ [/INST]$ as the template of input query and analysis instruction. The experiments are conducted on 4 NVIDIA 6000Ada GPUs. During training, the batch size per GPU is set to 4, and we use the AdamW optimizer with the learning rate and weight decay set to 1e-5 and 0.05, respectively. We also adopt linear Warmup-CosineLR as the learning rate scheduler, setting the warmup iterations to 1000. For image preprocessing, the image is resized to (448, 448) during training and testing and normalized with the mean and variant of [0.55,0.55,0.53] and [0.22,0.23,0.25].

## 4.2. Quantitative Results

In this section, we first compare the results of the proposed ANNEXE with other methods to demonstrate its effectiveness. Then, we conduct ablation experiments on the depth estimation branch to indicate the benefit of this auxiliary task for mask grounding. Finally, we set different hyperparameters for comparison to show the influences of various loss weights.

### 4.2.1. Comparison Results

To validate the effectiveness of the proposed ANNEXE model in understanding the egocentric interactions, we conduct comparison results with other methods on the Ego-IRGBench dataset. However, to the best of our knowledge, few methods can accomplish analyzing, answering, and pixel grounding in a unified manner. Therefore, we compare our method with other similar methods that can achieve partial functionality. Specifically, we first compare the interaction analyzing and answering abilities of the ANNEXE with some image captioning methods, which can generate descriptions of happening interactions according to instructions and queries. Furthermore, we compare the pixel-level grounding ability of ANNEXE with some referring image segmentation methods, which can generate mask grounding results based on the query.

**Interaction analyzing and answering results.** The first and second sub-tasks of the proposed Ego-IRG task are to generate descriptions of interactions between hands and objects and answer what needs to be segmented according to the query, which is similar to the image captioning task given different text prompts. Therefore, we selected some models in image captioning to validate the effectiveness of the interaction analyzing and answering abilities of ANNEXE. The comparison results of analyzing and answering results are shown in Tab. 2 and Tab. 3, respectively. Our proposed ANNEXE model achieves the best results even compared with other LLM-based methods such as Small-

Table 2. ANNEXE versus other methods about analyzing sub-task on the Ego-IRGBench validation and test sets. The metrics used are METEOR and CIDEr.

| Model | Val | | Test | |
|---|---|---|---|---|
| | M ↑ | C ↑ | M ↑ | C ↑ |
| ViECap (ICCV 2023) [17] | 0.063 | 0.096 | 0.096 | 0.094 |
| PromptCap (ICCV 2023) [20] | 0.147 | 0.176 | 0.148 | 0.184 |
| SmallCap (CVPR 2023) [58] | 0.177 | 0.219 | 0.179 | 0.221 |
| ConZIC (CVPR 2023) [81] | 0.102 | 0.063 | 0.102 | 0.067 |
| **Ego-IRG** | **0.563** | **1.516** | **0.563** | **1.494** |

Table 3. ANNEXE versus other methods about answering sub-task on the Ego-IRGBench validation and test sets. The metrics used are METEOR and CIDEr.

| Model | Val | | Test | |
|---|---|---|---|---|
| | M ↑ | C ↑ | M ↑ | C ↑ |
| ViECap (ICCV 2023) [17] | 0.019 | 0.024 | 0.021 | 0.026 |
| SmallCap (CVPR 2023) [58] | 0.034 | 0.091 | 0.035 | 0.092 |
| PromptCap (ICCV 2023) [20] | 0.086 | 0.205 | 0.089 | 0.218 |
| **Ego-IRG** | **0.363** | **2.543** | **0.365** | **2.590** |

Table 4. ANNEXE versus other referring image segmentation methods on the Ego-IRGBench validation and test sets to prove the pixel grounding capability. The metric for pixel-grounding is cIoU.

| Model | Val.cIoU ↑ | Test.cIoU ↑ |
|---|---|---|
| GSVA (CVPR 2024) [74] | 12.92 | 13.32 |
| GRES (CVPR 2023) [44] | 19.88 | 21.35 |
| LISA (CVPR 2024) [28] | 21.74 | 22.63 |
| X-Decoder (CVPR 2023) [86] | 23.72 | 23.75 |
| **Ego-IRG** | **35.14** (+11.42) | **36.02** (+12.27) |

Cap [58] and PromptCap [20], which can generate fluent and coherent descriptions and answers about the egocentric interactions taking place according to the query.

**Pixel grounding results.** To validate the performance of the pixel grounding of ANNEXE, we compare it with some of the latest referring image segmentation methods (including LLM-based methods such as GSVA [74] and LISA [28]) on the validation and test sets of the Ego-IRGBench dataset. The results are shown in Tab. 4, which shows that our ANNEXE achieves state-of-the-art segmentation performance in referring the objects involving interactions according to queries. Therefore, the proposed ANNEXE model can efficiently understand the context of multi-modal inputs and perform fine-grained pixel grounding for various downstream tasks.

### 4.2.2. Ablation Study

As introduced in Sec. 3.2.2, we integrate depth estimation as an auxiliary task for precise pixel grounding. In this section, we conduct the ablation study for the depth estimation branch to prove its ability to distinguish the foreground and background. The results are shown in Tab. 5. As observed

Table 5. Ablation study about the depth estimation branch.

| | Sub-task | Metrics | Without depth | With depth |
|---|---|---|---|---|
| Val | Analyzing | METEOR | 0.537 | 0.563 |
| | | CIDEr | 1.447 | 1.516 |
| | Answering | METEOR | 0.352 | 0.363 |
| | | CIDEr | 2.312 | 2.543 |
| | Grounding | cIoU | 31.72 | 35.14 |
| Test | Analyzing | METEOR | 0.536 | 0.563 |
| | | CIDEr | 1.423 | 1.494 |
| | Answering | METEOR | 0.350 | 0.365 |
| | | CIDEr | 2.317 | 2.590 |
| | Grounding | cIoU | 31.90 | 36.02 |

Table 6. Hyper-parameter experiments of ANNEXE on the Ego-IRGBench test set. 'Ans.' means answering sub-task, 'Ana.' means the analyzing sub-task, 'M' and 'C' represent the metrics METEOR and CIDEr, and cIoU is the metric of pixel grounding.

| $\lambda_{des}$ | $\lambda_{ans}$ | $\lambda_{pxl}$ | $\lambda_{cls}$ | $\lambda_{dep}$ | Ana.M | Ana.C | Ans.M | Ans.C | cIoU |
|---|---|---|---|---|---|---|---|---|---|
| 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.538 | 1.421 | 0.346 | 2.328 | 23.34 |
| 0.5 | 0.5 | 0.5 | 1.0 | 1.0 | 0.523 | 1.410 | 0.352 | 2.316 | 22.46 |
| **1.0** | **1.0** | **1.0** | **0.5** | **0.5** | **0.563** | **1.494** | **0.365** | **2.590** | **36.02** |

from the results, when the depth estimation is incorporated, the performance of pixel grounding will be improved by 3.42% and 4.12% on validation and test sets, respectively, which indicates the necessity of depth information.

### 4.2.3. Hyper-parameter Study

As described in sec. 3.2.3, the weights $\lambda_{des}$, $\lambda_{ans}$, $\lambda_{cls}$, $\lambda_{dep}$, and $\lambda_{pxl}$ are needed to be set during training. In this section, we study the influences of different hyper-parameter settings. The results are shown in Tab. 6, which exhibits great differences between diverse settings. Specifically, when the weights of analyzing, answering, and pixel grounding are smaller than the weights of indicator classification and depth estimation, the performance will be decreased. It is mainly because the interaction analyzing, answering, and pixel grounding are the primary goals of the proposed ANNEXE, and the depth estimation and index classification are designed to assist pixel grounding.

### 4.3. Qualitative Evaluation

Due to the page limitation, we show qualitative results for Ego-IRG task on the Ego-IRGBench dataset in Appendix A.4.

## 5. Conclusion

To comprehensively interpret egocentric interactions, this paper proposes the novel Ego-IRG task, which encompasses analyzing, answering, and pixel grounding sub-tasks. Furthermore, this paper creates a large-scale Ego-IRGBench dataset, which includes egocentric images and queries with corresponding multi-modal responses. Finally, the ANNEXE model is established to generate the text- and pixel-level outputs precisely according to diverse queries.

# 6. Acknowledgement

## References

[1] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022. 3

[2] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 3

[3] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1949–1957, 2015. 3

[4] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 7

[5] Siddhant Bansal, Michael Wray, and Dima Damen. Hoi-ref: Hand-object interaction referral in egocentric vision. *arXiv preprint arXiv:2404.09933*, 2024. 3

[6] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 3

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, pages 1877–1901, 2020. 3

[8] Tejo Chalasani, Jan Ondrej, and Aljosa Smolic. Egocentric gesture recognition for head-mounted ar devices. In *IEEE international symposium on mixed and augmented reality adjunct*, pages 109–114. IEEE, 2018. 1

[9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 4, 5, 7

[10] Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *arXiv preprint arXiv:2408.14469*, 2024. 3

[11] Yifei Chen, Dapeng Chen, Ruijin Liu, Sai Zhou, Wenyuan Xue, and Wei Peng. Align before adapt: Leveraging entity-to-region alignments for generalizable video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18688–18698, 2024. 1

[12] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024. 3

[13] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, 130:33–55, 2022. 2, 6

[14] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 2, 6

[15] Shangzhe Di and Weidi Xie. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943, 2024. 2, 3

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 4

[17] Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146, 2023. 8

[18] Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang Wang, and Rakesh Ranjan. Mmg-ego4d: Multimodal generalization in egocentric action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6481–6491, 2023. 3

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 2, 6

[20] Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the*

*IEEE/CVF International Conference on Computer Vision*, pages 2963–2975, 2023. 8

[21] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *Proceedings of the International conference on machine learning*, pages 9118–9147. PMLR, 2022. 3

[22] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing*, 29:7795–7806, 2020. 3

[23] Baoxiong Jia, Ting Lei, Song-Chun Zhu, and Siyuan Huang. Egotaskqa: Understanding human tasks in egocentric videos. *Advances in Neural Information Processing Systems*, 35:3343–3360, 2022. 3

[24] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 6

[25] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *Proceedings of the European Conference on Computer Vision*, pages 140–158. Springer, 2025. 3

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6

[27] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 3

[28] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 3, 7, 8

[29] Kyungjun Lee, Abhinav Shrivastava, and Hernisa Kacorri. Hand-priming in object localization for assistive egocentric vision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3422–3432, 2020. 3

[30] Ting Lei, Shaofeng Yin, and Yang Liu. Exploring the potential of large foundation models for open-vocabulary HOI detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16657–16667. IEEE, 2024. 1

[31] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Egocentric human-object interaction detection exploiting synthetic data. In *International Conference on Image Analysis and Processing*, pages 237–248. Springer, 2022. 3

[32] Rosario Leonardi, Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario. *Computer Vision and Image Understanding*, 242:103984, 2024. 1

[33] Rosario Leonardi, Antonino Furnari, Francesco Ragusa, and Giovanni Maria Farinella. Are synthetic data useful for egocentric hand-object interaction detection? In *Proceedings of the European Conference on Computer Vision*, pages 36–54. Springer, 2025. 2

[34] Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14497–14509, 2024. 1

[35] Haoxin Li, Yijun Cai, and Wei-Shi Zheng. Deep dual relation modeling for egocentric interaction recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7932–7941, 2019. 3

[36] Haoxin Li, Wei-Shi Zheng, Jianguo Zhang, Haifeng Hu, Jiwen Lu, and Jian-Huang Lai. Egocentric action recognition by automatic relation modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):489–507, 2022. 2

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[38] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 3

[39] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[40] Shuang Li, Xavier Puig, Chris Paxton, Yilun Du, Clinton Wang, Linxi Fan, Tao Chen, De-An Huang, Ekin Akyürek, Anima Anandkumar, et al. Pre-trained language models for interactive decision-making. *Advances in Neural Information Processing Systems*, 35:31199–31212, 2022. 3

[41] Yin Li, Zhefan Ye, and James M Rehg. Delving into egocentric actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 287–295, 2015. 3

[42] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European conference on computer vision*, pages 619–635, 2018. 6

[43] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *Advances in Neural Information Processing Systems*, 35:7575–7586, 2022. 3

[44] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings*

*of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23592–23601, 2023. 2, 8

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[46] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 704–721. Springer, 2020. 3

[47] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Jun Kong. Enhanced attention tracking with multi-branch network for egocentric activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(6):3587–3602, 2021. 3

[48] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022, 2022. 2, 6, 7

[49] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[50] Yao Lu and Walterio W Mayol-Cuevas. Egocentric hand-object interaction detection and application. *arXiv preprint arXiv:2109.14734*, 2021. 3

[51] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 6

[52] Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 3

[53] Shan Ning, Longtian Qiu, Yongfei Liu, and Xuming He. Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23507–23517, 2023. 1

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 3

[55] Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18286–18296, 2024. 3

[56] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[58] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 8

[59] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 3, 6, 7

[60] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 3

[61] Yang Shen, Bingbing Ni, Zefan Li, and Ning Zhuang. Egocentric activity prediction via event modulated attention. In *Proceedings of the European conference on computer vision*, pages 197–212, 2018. 3

[62] Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torresani, and Effrosyni Mavroudi. Learning to segment referred objects from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14510–14520, 2024. 2, 6

[63] Tsukasa Shiota, Motohiro Takagi, Kaori Kumagai, Hitoshi Seshimo, and Yushi Aono. Egocentric action recognition by capturing hand-object contact and object state. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6541–6551, 2024. 3

[64] Yuejiao Su, Yi Wang, and Lap-Pui Chau. Ormnet: Object-centric relationship modeling for egocentric hand-object segmentation. *arXiv preprint arXiv:2407.05576*, 2024. 2, 3

[65] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9954–9963, 2019. 3

[66] Alessandro Suglia, Claudio Greco, Katie Baker, Jose L Part, Ioannis Papaioannou, Arash Eshghi, Ioannis Konstas, and Oliver Lemon. Alanavlm: A multimodal embodied ai foundation model for egocentric video understanding. *arXiv preprint arXiv:2406.13807*, 2024. 1

[67] Hector A Valdez, Kyle Min, and Subarna Tripathi. Svitt-ego:

A sparse video-text transformer for egocentric video. *arXiv preprint arXiv:2406.09462*, 2024. 3

[68] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7

[69] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*, 2023. 3

[70] Huiyu Wang, Mitesh Kumar Singh, and Lorenzo Torresani. Ego-only: Egocentric action detection without exocentric transferring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5250–5261, 2023. 2

[71] Xiaohan Wang, Yu Wu, Linchao Zhu, and Yi Yang. Symbiotic attention with privileged information for egocentric action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12249–12256, 2020. 3

[72] Xiaohan Wang, Linchao Zhu, Yu Wu, and Yi Yang. Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE transactions on pattern analysis and machine intelligence*, 45(6):6605–6617, 2020. 3

[73] Xiaohan Wang, Linchao Zhu, Heng Wang, and Yi Yang. Interactive prototype learning for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 3

[74] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3858–3869, 2024. 2, 7, 8

[75] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 2

[76] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 5, 7

[77] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 3

[78] Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90, 2024. 3

[79] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2155–2163, 2023. 3

[80] Olga Zatsarynna, Yazan Abu Farha, and Juergen Gall. Multi-modal temporal convolutional network for anticipating actions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2249–2258, 2021. 3

[81] Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476, 2023. 8

[82] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pevl: Pose-enhanced vision-language model for fine-grained human action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18867, 2024. 1

[83] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *Proceedings of the European Conference on Computer Vision*, pages 127–145. Springer, 2022. 2, 3, 6

[84] Yang Zhou, Bingbing Ni, Richang Hong, Xiaokang Yang, and Qi Tian. Cascaded interactional targeting network for egocentric video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1904–1913, 2016. 3

[85] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 6

[86] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15116–15127, 2023. 8