

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

Supplementary Material

Description: In the image, the right hand of the person is holding a black laptop.



Egocentric RGB Image



Egocentric Depth Map

Query 1: What is held by hand?



Answer 1: Th **laptop** is held by hand, and the mask of it is **<MSK>**.

Example 1

Query n: The hands involving interactions.



Answer n: The laptop is held by **the right hand**, and the mask of the hand is **<MSK>**.

Example n

Figure 1. **Structure of the Ego-IRGBench dataset.** For each egocentric image, a depth map and a query about interacting are provided. In addition, the query-answer-mask annotations are also included.

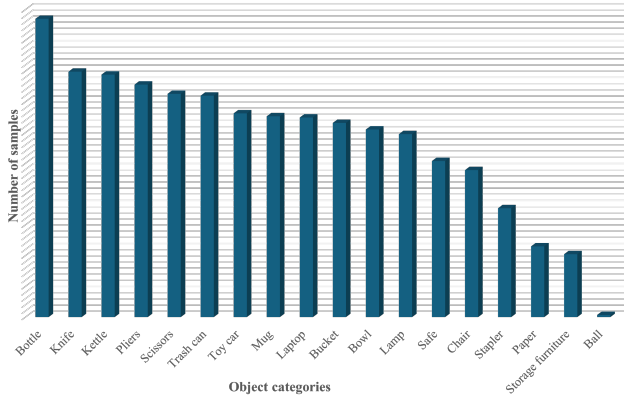


Figure 2. Distribution of interacting object categories in our Ego-IRGBench dataset.

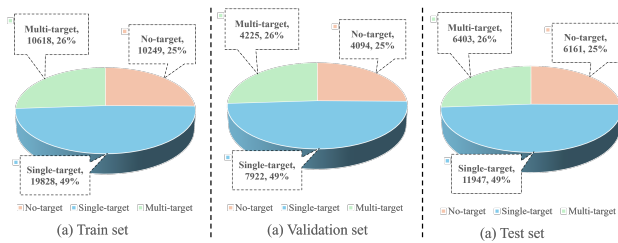


Figure 3. The proportion of single-target, multi-target, and no-target samples in the Ego-IRGBench training, validation, and test sets.

In this supplementary material, we first introduce the structure and scale of the proposed Ego-IRGBench dataset in Sec. A.1. The semi-automatic step-wise annotation pipeline is introduced in Sec. A.2. To build the benchmark, we explain the detailed criteria used in the Ego-IRG task in Sec. A.3. Finally, the visualization of the experimental results of the ANNEXE model is shown in Sec. A.4.

A. Ego-IRGBench

A.1. Structure and Scale

[4] In this paper, we create a large-scale dataset Ego-IRGBench for Ego-IRG (Egocentric reasoning and pixel grounding) task, which includes 20,681 RGB-D egocentric images and over 1.6 million queries about the interactions, along with corresponding textual and pixel-level responses. The structure of the dataset is shown in Fig. 1.

We calculate the distribution of object categories in our dataset, which is shown in Fig. 2. We can observe that the Ego-IRGBench dataset includes 18 object categories. Some hard categories are also included with a limited number of samples, such as “ball.” In addition, diverse queries for no-target, single-target, and multi-target are included in our Ego-IRGBench dataset. Thus, we count various types of queries on train, validation, and test sets, respectively, as depicted in Fig. 3. We can observe that the proportion of different types of queries is basically consistent, which indicates the reasonableness of the dataset allocation.

A.2. Step-wise Annotation Pipeline

To establish the Ego-IRGBench dataset, we employed experts to relabel the dataset based on the original HOI4D [3] dataset. In specific, we employed five graduate students with a master’s degree or above in artificial intelligence as experts to accurately label the data, and each person spent nearly 30 hours. Furthermore, We developed an effective semi-automatic step-wise annotation pipeline to decrease the manual efforts consumed in building the dataset. The pipeline is shown in Fig. 4. The whole pipeline includes three essential steps: interaction classification, hand-object mask generation, and query-response generation.

Step 1: interaction classification. The HOI4D [3] dataset is a 4D egocentric dataset for category-level

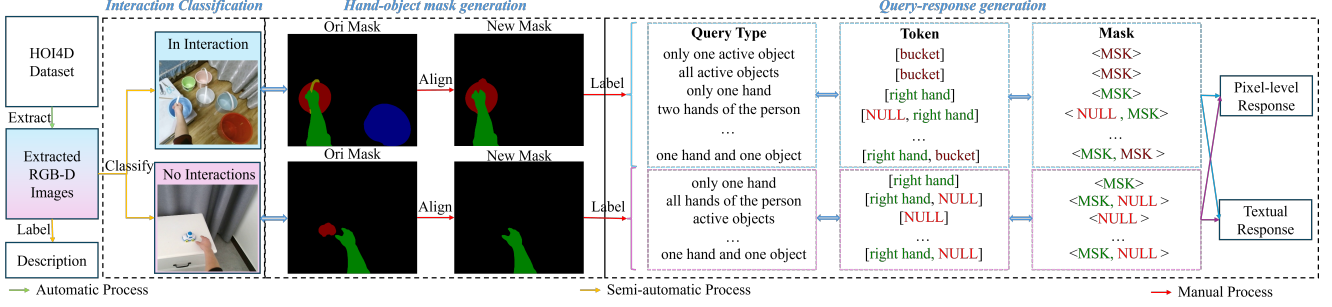


Figure 4. **The overall pipeline to re-label the dataset for the Ego-IRG task.** The pipeline includes three steps: interaction classification, hand-object mask generation, and query-response generation. The interaction classification step is to determine whether interaction takes place, the hand-object mask generation step is to align the original mask to our categories, and the query-response generation step is to label the final textual and pixel-level response for each query.

human-object interaction, which includes egocentric RGB-D videos and corresponding action categories, motion segmentation masks, 3D hand and object poses, object meshes, *etc.* Built upon the HOI4D [3] dataset, we first extracted RGB-D frames according to different actions at equal time intervals from egocentric videos with corresponding segmentation masks (original mask). Furthermore, to establish the dataset efficiently, we adopt the interaction classification first, which is to classify the extracted RGB-D pairs into two categories: “In Interaction” and “No Interactions.” For images where interaction occurs, we use a semi-manual approach to label the extracted frames with description annotations. Specifically, we manually labeled 400 images with descriptions of interactions between hands and objects. Then, we used these descriptions to fine-tune the multi-modal large language model [6] to generate more descriptions. Finally, the generated descriptions are checked and modified by experts to be more diverse and comprehensive. For images without interactions, we generated a description set including {“The <left hand> is not interacting with anything”, “The person is hanging out the <hands> without any interactions,”...}. The final descriptions of these images are randomly selected from the description set.

Step 2: hand-object mask generation. The extracted RGB-D egocentric frames are equipped with masks (original masks) in the HOI4D [3] dataset. However, the original masks cannot fulfill the requirements of our Ego-IRG task. Therefore, we asked the experts to align the original mask to generate the new mask manually. Specifically, we set up five categories: left hand, right hand, objects interacting with the left hand, objects interacting with the right hand, and objects interacting with both hands. During the annotation process, we require that only masks of these five categories be saved or annotated, and other irrelevant categories in the original mask are deleted.

Step 3: query-response generation. After generating the masks in step 2, we generated the query and corresponding textual and pixel-level responses manually. Specifically,


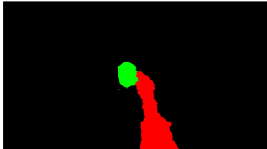
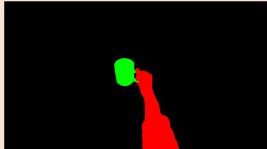
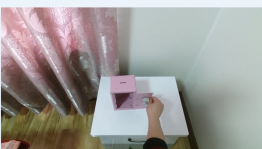

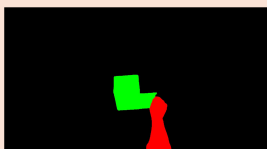

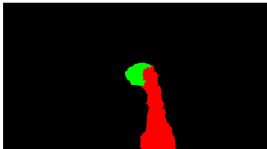
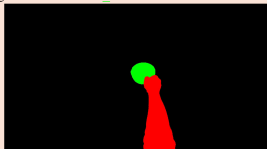
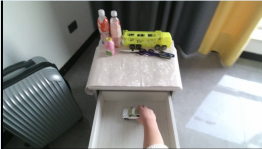


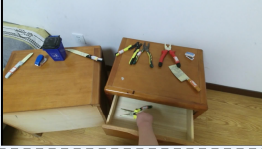
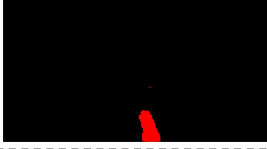
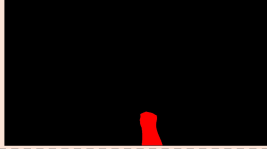



we disassemble and combine the masks annotated in the second step, annotate different queries and tokens according to different mask combinations, and obtain the corresponding query-token and mask pairs. The token is the basic component of the answer, which is the name of the objects that need to be segmented. In addition, we used a template to generate the answers according to the mask and token. The template we used is “The mask of < $token_1$ > is < MSK_1 >, the mask of < $token_2$ > is < MSK_2 >, ..., and the mask of < $token_n$ > is < MSK_n >”. In this way, for each egocentric image, the query and corresponding text- and pixel-level responses are generated comprehensively.

A.3. Criteria

To build up the benchmark for the Ego-IRG task, we set up the detailed evaluation criteria for three sub-tasks. Specifically, for the analyzing sub-task, which aims to generate the descriptions of interactions between hands and objects, we evaluate the quality of generated descriptions using METEOR [1] and CIDEr [5] metrics. Also, these two metrics are used to evaluate the answering sub-task, which aims to answer the query in textual response. In addition, for the pixel grounding sub-task, we use cIoU [2] to verify the accuracy of generated masks according to queries regarding interactions.

A.4. Experimental Results

We visualize some of the experimental results of our AN-NEXE in this section, the results of which are shown in Fig. 5. The results of the multi-target samples are shown in the first to third rows, and the results of single-target and no-target samples are exhibited in the fourth and fifth rows. We can observe that the model can generate fluent and accurate descriptions for each egocentric image, which is semantically analogous to the description of ground truth (GT). The predicted answers for the query are also coherent and natural, along with precise pixel-level mask responses for queries regarding interactions.

<p>Input Query: Please segment the hands and interacting objects.</p> 	<p>Description Prediction: This image shows a person using the right hand to make contact with a mug. Answer Prediction: The mask of the right hand is <MSK_0> and the mask of the active mug interacting with right hand is <MSK_1>.</p> 	<p>Description GT: The picture shows a person's right hand grasping a mug. Answer GT: The mask of the right hand is <MSK_0> and the mask of the active mug interacting with right hand is <MSK_1>.</p> 
<p>Input Query: Please segment the hands and interacting objects.</p> 	<p>Description Prediction: The image image represents a person manipulating a safe using the right hand. Answer Prediction: The mask of the right hand is <MSK_0> and the mask of the active safe interacting with right hand is <MSK_1>.</p> 	<p>Description GT: This image captures a person operating a safe with the right hand. Answer GT: The mask of the right hand is <MSK_0> and the mask of the active safe interacting with right hand is <MSK_1>.</p> 
<p>Input Query: Hands and objects in interaction.</p> 	<p>Description Prediction: This image shows a person holding a bowl in the right hand. Answer Prediction: The mask of the right hand is <MSK_0> and the mask of the active bowl interacting with right hand is <MSK_1>.</p> 	<p>Description GT: This image shows a person using the right hand to utilize a bowl. Answer GT: The mask of the right hand is <MSK_0> and the mask of the active bowl interacting with right hand is <MSK_1>.</p> 
<p>Input Query: The hand that is not the left hand of the person.</p> 	<p>Description Prediction: This image shows a person standing in front of the table and interacting with a small toy car with the right hand. Answer Prediction: The mask of the right hand is <MSK_0>.</p> 	<p>Description GT: This image depicts an individual gripping a toy car with the right hand. Answer GT: The mask of the right hand is <MSK_0>.</p> 
<p>Input Query: Please segment the hands of the camera wearer.</p> 	<p>Description Prediction: This image image shows a person standing in front of the table and wielding a pair of pliers with the right hand Answer Prediction: The mask of the right hand is <MSK_0>.</p> 	<p>Description GT: The image depicts the camera wearer standing in front of the table, with the right hand holding a pair of pliers Answer GT: The mask of the right hand is <MSK_0>.</p> 
<p>Input Query: Please segment the lamp interacting with left hand and right hand.</p> 	<p>Description Prediction: This image shows a person using the right hand to approach a lamp. Answer Prediction: There is no suitable results.</p> 	<p>Description GT: The image displays someone with the right hand held out, not making contact with any items Answer GT: There is no suitable results.</p> 

(a) Egocentric Image and Query

(b) Predicted Description, Answer, and Mask

(c) Description, Answer, and Mask Ground Truth

Figure 5. Qualitative visualization results of our ANNEXE model on Ego-IRGBench validation and test sets.

References

- [1] Satyanjeev 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. [2](#)
- [2] 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. [2](#)
- [3] 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. [1](#), [2](#)
- [4] Shentong Mo and Yapeng Tian. Multi-modal grouping network for weakly-supervised audio-visual video parsing. In *Advances in Neural Information Processing Systems*, 2022. [1](#)
- [5] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. [2](#)
- [6] 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. [2](#)