Learning from Observer Gaze: Zero-Shot Attention Prediction Oriented by Human-Object Interaction Recognition Supplementary Materials

Yuchen Zhou, Linkai Liu, Chao Gou*
Sun Yat-sen University
https://yuchen2199.github.io/Interactive-Gaze/
{zhouych37,liulk6}@mail2.sysu.edu.cn, gouchao@mail.sysu.edu.cn

Abstract

In the supplementary materials, we provide further details about the proposed Interactive Gaze dataset, including the data sources, the gaze data collection procedure, and more visualization and analysis (Sec. 1). We also include more experimental details (Sec. 2) and additional experimental results (Sec. 3).

1. Details about Interactive Gaze Dataset

1.1. Data source

The images utilized in Interactive Gaze (IG) are derived from VCOCO [3] and HICO-det [1] datasets. Specifically, we selected 4475 samples from the VCOCO training set, 720 samples from the VCOCO test set, and 1104 samples from the HICO-det test set. Our initial sampling involved a random selection process, followed by the exclusion of samples featuring ambiguous images, unclear semantics, or incorrect labeling. Subsequently, to address the categories overlooked by the initial random sampling, additional samples were deliberately included. As a result of this refined sampling strategy, the ensuing interactions exhibit an average of 13 samples per category. Notably, the <*person, sit on, chair*> category boasts the maximum representation with 91 samples, while the <*person, carry, clock*> category has a minimal presence, with just one sample. The limited number of samples in certain categories can be attributed to their inherent scarcity within the VCOCO and HICO-det.

1.2. Gaze data collection procedure

We employ a mouse-click-based paradigm for acquiring gaze fixation data, a methodology akin to previous research [2, 4, 6, 7] grounded in neurophysiological and psychophysical studies, which has demonstrated widespread efficacy. This program involves the development of a human-computer interaction procedure designed to emulate human gaze patterns through mouse clicks. It facilitates large-scale data collection on gaze fixation by capturing the behavior through a generic mouse, obviating the need for a dedicated eye tracker.

A total of 32 participants, comprising 18 males and 14 females aged between 19 and 29 years, and possessing either normal or corrected-to-normal vision, took part in this study. The study was approved by the institutional review board. Before the commencement of the experiment, each participant thoroughly read and signed the informed consent form. Additionally, they underwent a pre-experiment orientation to familiarize themselves with the procedures. Monetary compensation was provided upon completion of the experiment.

During the formal experiment, participants were presented with visual images depicting a specific person and object engaged in an interaction, accompanied by text describing the interaction. Subsequently, participants were tasked with identifying the key visual cues within the images most pertinent to the provided interaction. It is noteworthy that participants were specifically instructed not to focus on cues for recognizing persons and objects, as previous gaze fixation datasets have been extensively explored. Instead, their attention was directed toward key visual cues associated with the interaction being performed.

To allow participants ample time to comprehend the image content, they were given control over the switching of experimental samples, with a minimum limit of 6 seconds per sample. All image samples were standardized to a uniform size of 400×400 pixels for normalization purposes, and the order of presentation was randomized. Participants were permitted breaks during the experiment to ensure a sustained and optimal mental state for subsequent tasks.

Upon concluding all experiments, we aggregated and applied a Gaussian filter to blur all mouse-click-like data as-

*Corresponding author.

sociated with the same sample, thereby generating salient maps, following previous work [4, 8].

1.3. More visualization and analysis

Data statistics. Figure 1 illustrates the 60 most frequently occurring interaction (HOI) categories in the proposed Interactive Gaze dataset. Additionally, Figure 2 depicts the 30 most frequently occurring object categories, while Figure 3 highlights the 30 most frequently occurring verb categories within the same Interactive Gaze dataset.

Category organization. A detailed hierarchy HOI category organization of the proposed Interactive Gaze dataset is shown in Figure 4.

Comparison of mean fixation heatmaps. We compare the mean fixation heatmaps of the object-centered representative dataset SALICON [4] and our proposed interaction-centered IG dataset. As shown in Figure 5, both datasets exhibit center biases. However, SALICON displays a more pronounced bias, while the fixation probability of the IG dataset remains relatively dispersed. This also emphasizes that our proposed prediction of interaction-oriented attention is significantly more challenging compared to previous object-centered attention predictions.

Visualization cases. Figure 6 presents visualizations of the 48 instances from the proposed Interactive Gaze dataset. In the first row, instances involve the same verb *"carry"* but with different objects. The second row displays instances attributed to the same class of HOIs, i.e., <*person, cut, with scissors*>. The third and fourth rows illustrate interaction-oriented attention in sports and restaurant scenes, respectively. The fifth row showcases interaction-oriented attention in office and outdoor scenes. The sixth row focuses on visual attention corresponding to interactions with animals. Finally, the seventh and eighth rows complement the other instances.

2. Experimental Details

ZeroIA setting. In light of the inherent diversity and nearly boundless nuances within interactions, we introduce the Zero-shot Interaction-oriented Attention (ZeroIA) prediction task. This task is designed to assess the model's proficiency in efficiently recognizing previously unseen interaction categories. Within the ZeroIA setting, we designate the 213 human-object interaction (HOI) classes present in the IG dataset, originating from the VCOCO training set, as seen classes. The attention prediction model is trained using these classes. Concurrently, we set the 527 HOI classes in the VCOCO test set and the HICO test set, which do not overlap with the VCOCO training set, as unseen classes. These unseen classes serve as the benchmark for evaluating the attention prediction model and contain a total of 1,105 samples.

Fully supervised setting. In the fully supervised setting, we train attention prediction models with the 213 HOI classes in the IG dataset originating from the VCOCO training set and test them with 719 samples in the IG dataset originating from the VCOCO test set.

3. More Experimental Results

Visualization of diverse model variants. Here, we sequentially exclude the positional adapter (PA), the visual adapter (VA), the human-object cognitive block (HOCB), and the interaction cognitive block (ICB) to create variant models of the proposed Interactive Attention model. The corresponding predicted results of these variant models are shown in Figure 7.

Predicted results and attention visualizations of the HOI method. We present more visualizations of the cross-attention maps of interaction branches in original MUREN [5] and the MUREN aligned with interaction-oriented attention, respectively, as shown in Figure 8. It is evident that the attention map of the original MUREN appears fragmented and struggles to focus on interaction-related visual cues, leading to failures in interaction recognition. Conversely, after aligning with interaction-oriented attention, not only are the erroneous results corrected, but the attention map becomes significantly more interpretable and focuses on key regions.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, pages 381–389. IEEE, 2018. 1

[2] Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *NIPS*, 35:9432–9446, 2022. 1

[3] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 1

[4] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 1, 2, 6

[5] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, pages 2925–2934, 2023. 2

[6] Drew Linsley, Sven Eberhardt, Tarun Sharma, Pankaj Gupta, and Thomas Serre. What are the visual features underlying human versus machine vision? In *ICCVW*, pages 2706–2714, 2017. 1

[7] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *ICLR*. OpenReview.net, 2019. 1

[8] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 2

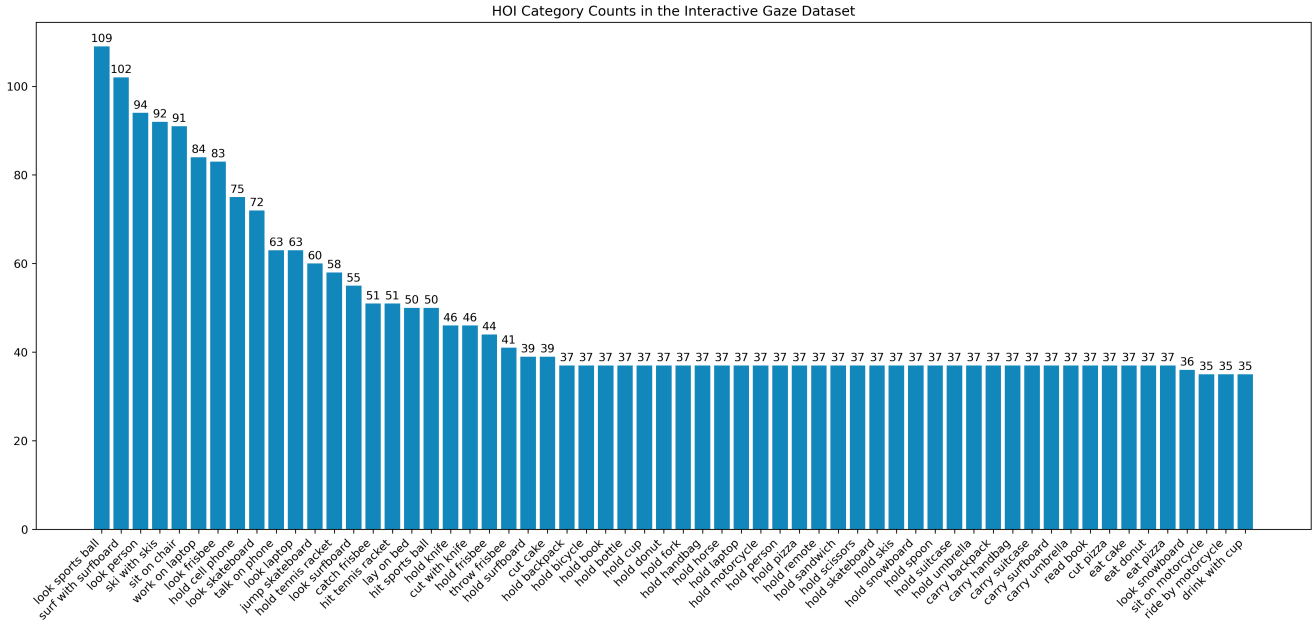

Figure 1. 60 most frequently occurring interaction (HOI) categories in the proposed Interactive Gaze dataset.

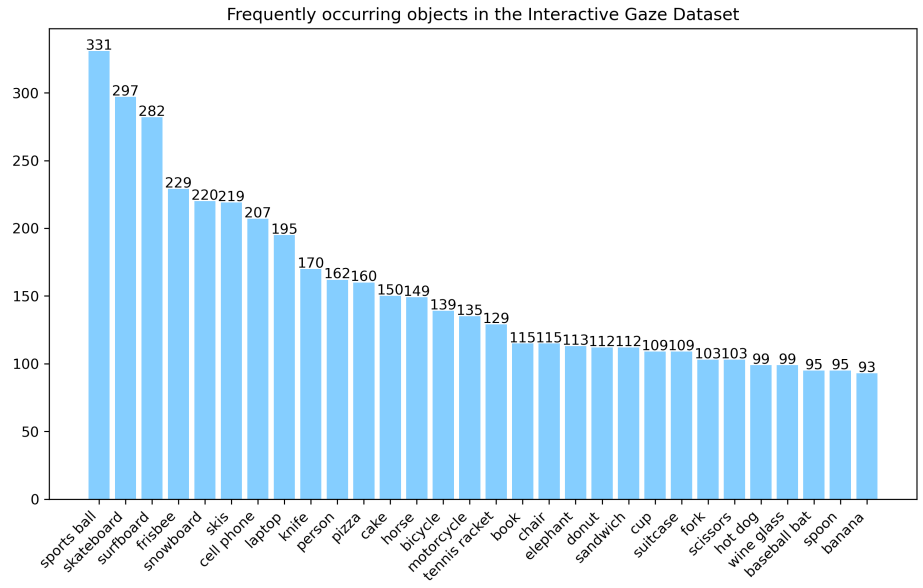

Figure 2. 30 most frequently occurring object categories in the proposed Interactive Gaze dataset.

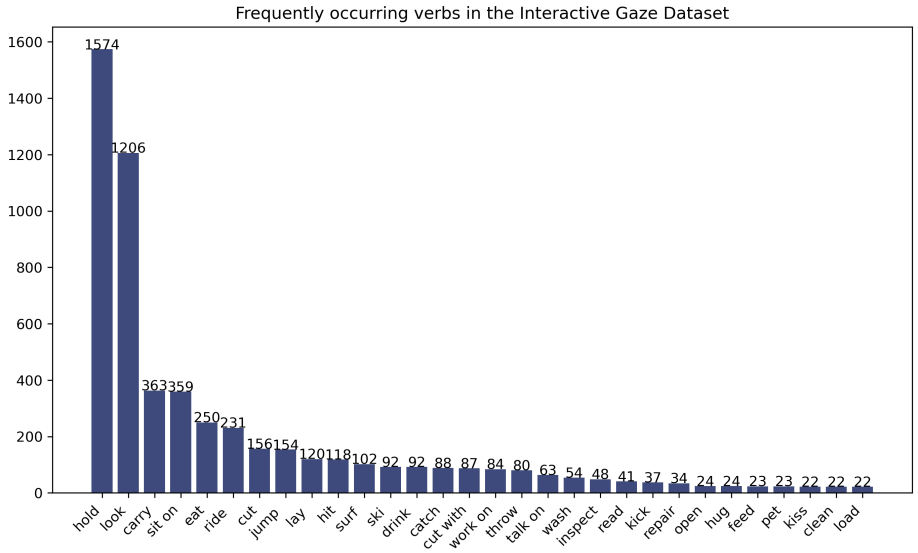

Figure 3. 30 most frequently occurring verb categories in the proposed Interactive Gaze dataset.

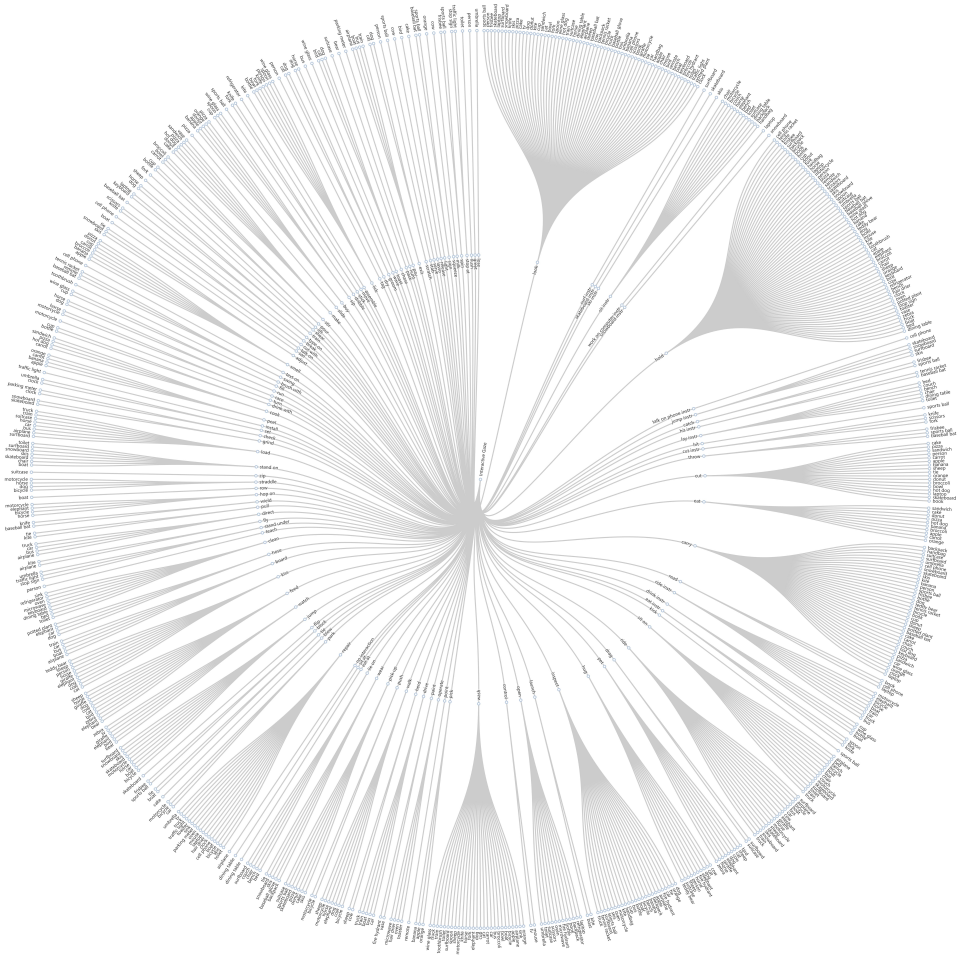

Figure 4. A detailed visualization of hierarchy HOI category organization in the proposed Interactive Gaze dataset. (zoom in for detail)

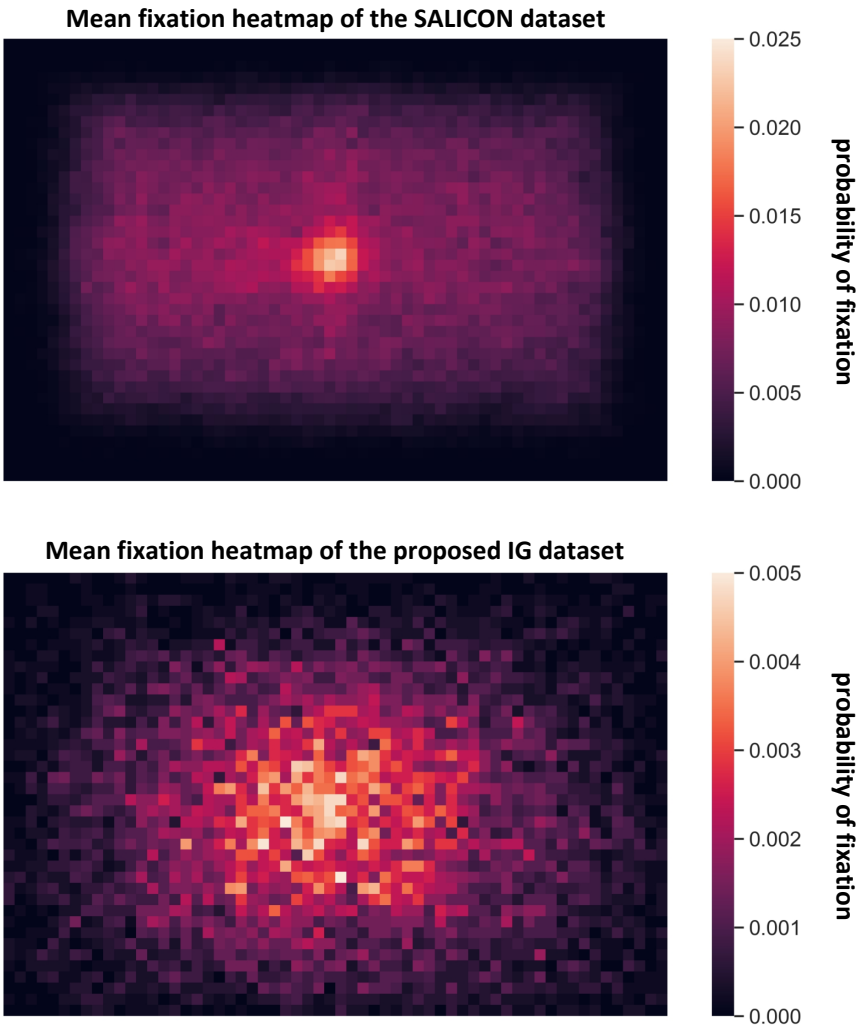

Figure 5. The mean fixation heatmaps of the object-centered representative dataset SALICON [4] and our proposed interaction-centered IG dataset. Both datasets demonstrate center biases, but SALICON exhibits a more pronounced bias, whereas the fixation probability of the IG dataset remains relatively dispersed. This also emphasizes that our proposed prediction of interaction-oriented attention is significantly more challenging compared to previous object-centered attention predictions.

Figure 6. More exemplars from the proposed Interactive Gaze dataset. This presents visualizations of the 48 instances. In the first row, instances involve the same verb *"carry"* but with different objects. The second row displays instances attributed to the same class of HOIs, i.e., *<person, cut with, scissors>*. The third and fourth rows illustrate interaction-oriented attention in sports and restaurant scenes, respectively. The fifth row showcases interaction-oriented attention in office and outdoor scenes. The sixth row focuses on visual attention corresponding to interactions with animals. Finally, the seventh and eighth rows complement the other instances. (zoom in for detail)

Figure 7. Visualization of diverse model variants. We sequentially exclude the positional adapter (PA), the visual adapter (VA), the human-object cognitive block (HOCB), and the interaction cognitive block (ICB) to create variant models of the proposed Interactive Attention model.

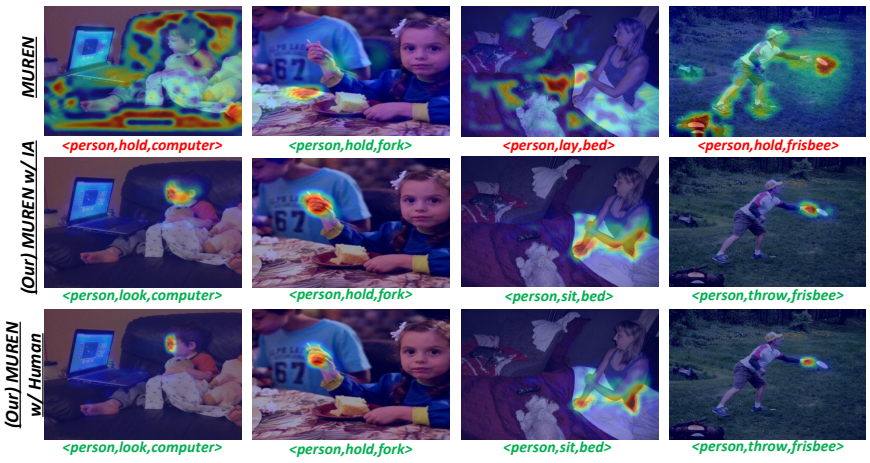

Figure 8. More predicted results and corresponding attention visualizations for MUREN and our (MUREN w/ IA, MUREN w/ Human). We mark true positive results in green, and false positive results in red. After aligning interaction-oriented attention, the erroneous prediction results are corrected, and the corresponding attention becomes more converged and more interpretable.