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

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

Most existing attention prediction research focuses on salient instances like humans and objects. However, the more complex interaction-oriented attention, arising from the comprehension of interactions between instances by human observers, remains largely unexplored. This is equally crucial for advancing human-machine interaction and human-centered artificial intelligence. To bridge this gap, we first collect a novel gaze fixation dataset named IG, comprising 530,000 fixation points across 740 diverse interaction categories, capturing visual attention during human observers' cognitive processes of interactions. Subsequently, we introduce the zero-shot interaction-oriented attention prediction task (ZeroIA), which challenges models to predict visual cues for interactions not encountered during training. Thirdly, we present the Interactive Attention model (IA), designed to emulate human observers' cognitive processes to tackle the ZeroIA problem. Extensive experiments demonstrate that the proposed IA outperforms other state-of-the-art approaches in both ZeroIA and fully supervised settings. Lastly, we endeavor to apply interaction-oriented attention to the interaction recognition task itself. Further experimental results demonstrate the promising potential to enhance the performance and interpretability of existing state-of-the-art HOI models by incorporating real human attention data from IG and attention labels generated by IA.

1. Introduction

Human visual attention reveals a wealth of information about internal cognitive states, including bottom-up visual stimulation and top-down goal-directed empirical knowledge [10,47,61]. Predicting visual attention provides a valuable understanding of how humans perceive, comprehend, and interact with the world, especially within the context of goal-directed attention [3–5, 39, 46, 50]. This holds profound implications for enhancing human-computer interaction [11, 25, 49], and contributing to fields such as assistance for visual impaired [15, 16], education [1, 51], and autonomous driving [13, 21, 43, 65].

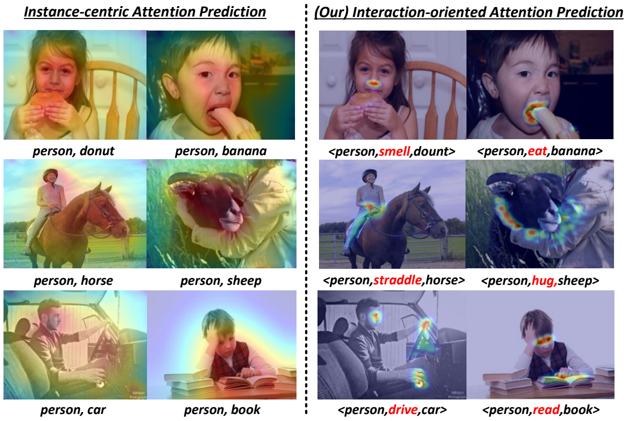

Figure 1. Previous attention prediction models have traditionally focused on the instance-level, primarily emphasizing foreground humans and objects. In contrast, our proposed interaction-oriented attention aims to capture subtler and more fine-grained visual cues associated with actions, such as body parts (row 1), human-object contact (row 2), and scene context (row 3). This proposition challenges the research community with a more intricate and cognitively demanding task.

Goal-oriented attention prediction studies have made some progress recently [8, 12, 24, 37, 42, 55–57], mostly focusing on predicting gaze fixation when searching or recognizing objects with relative explicit and invariant visual cues. However, interactions [27,38,68,69,71], another fundamental component of visual understanding that encapsulates how humans and objects interact with each other, remain largely unexplored. Interactions pose greater cognitive challenges than objects, due to their diverse and subtle visual cues, as shown in Figure 1. Consequently, predicting

[*] Corresponding author.

interaction-oriented visual attention proves to be a more demanding task compared to previous efforts. Moreover, the inherent diversity and almost infinite granularity of actions intersect with limited available data, necessitating the adoption of zero-shot learning. We name this problem *Zero*-shot *I*nteraction-oriented *A*ttention prediction (*ZeroIA*).

On the other hand, in the domain of human-object interaction (HOI) detection [29, 29, 34, 53, 60, 62, 64, 66, 67], current state-of-the-art methods also face limitations in the interaction comprehension phase, even though they demonstrate proficiency in handling the object detection phase. This limitation is rooted in the inherent difficulty of capturing visual cues related to actions.

To alleviate these problems, we first introduce *I*nteractive-*G*aze (*IG*), a novel dataset of gaze fixations capturing human cognitive process of interactions, filling a crucial gap in interaction-oriented visual attention research. IG contains 530,000 fixation points from 32 human observers, spanning 740 interaction categories, 80 objects, and 132 actions. Human observers are invited to capture key visual cues in various interaction scenes, while their visual attention during the cognitive process is recorded. All interaction scenes are selected from HOI benchmark datasets, HICO-det [2] and VCOCO [14]. This means that IG has the potential to bridge the fields of visual attention and HOI detection and be a catalyst for advancing both research areas simultaneously.

Secondly, inspired by the cognitive processes of human observers, we further propose a goal-oriented attention prediction approach, termed *I*nteractive *A*ttention (*IA*). Our approach begins by a set of clever interaction-oriented prompts and adapters to activate and leverage CLIP's powerful knowledge representation capabilities [48], thereby facilitating zero-shot learning. This process involves establishing adaptive knowledge prototypes based on each visual scene, avoiding direct retrieval of fixed knowledge from CLIP, which may exhibit strong noun bias [18, 41, 45]. Guided by these knowledge, IA first focuses on the perception and understanding of individual instances, i.e., humans and objects. Then, it further understands the interactions unfolding between instances, ultimately completing the generation of interaction-oriented attention. Extensive experiments demonstrate that IA outperforms state-of-the-art approaches in both ZeroIA and fully supervised settings.

Furthermore, we explore how goal-oriented attention can be fed back to the goal itself, particularly establishing an initial bidirectional pathway connecting goal-oriented attention and action understanding. We introduce a general and effective HOI training strategy. This can supplement most state-of-the-art model loss with additional supervision for interaction-oriented attention. Remarkably, extensive experiments reveal the substantial potential of visual attention for HOI models from two aspects. (1) Aligning a limited yet valuable genuine human interaction-oriented attention from the proposed IG enhances the performance and interpretability of existing state-of-the-art HOI models. (2) Integrating interaction-oriented attention generated by the proposed IA model further enhances HOI models, surpassing even the performance achieved in the context of human observers' attention. To summarize, the contributions of our work are three-fold:

- Firstly, we address a crucial gap in interaction-oriented attention studies by introducing the ZeroIA problem and IG dataset, marking a pioneering endeavor as the inaugural gaze fixation dataset specifically designed for interaction analysis. Given the composite nature of IG, it has significant potential to facilitate the intersection of multiple domains, such as goal-directed attention and interaction comprehension.
- Secondly, we present the IA model, a novel approach crafted to emulate human cognitive processes for predicting high-quality interaction-oriented attention. Extensive experiments demonstrate IA outperforms other state-of-the-art attention prediction methods in both the ZeroIA and fully-supervised settings.
- Thirdly, we introduce a general and simple HOI training strategy with attention and demonstrate that both aligning a small yet precious genuine human attention from IG and aligning attention generated by IA can enhance the performance and interpretability of existing state-of-the-art HOI models.

2. Interactive Gaze (IG) Dataset

To study human visual attention oriented by HOI, we collect the Interactive Gaze (IG), the first large-scale interaction-centric gaze fixation dataset, as shown in Figure 2. IG comprises 6,299 interaction scenarios across 740 interaction categories, 80 object categories, and 132 action categories. IG captures the visual attention of 32 human observers during the cognition of these interaction scenarios, resulting in 530,000 corresponding fixation points. IG holds substantial potential to bridge the domains of visual attention and action understanding, serving as a catalyst to jointly promote these two areas of study. We use mouse clicks to simulate gaze fixation for visual attention studies [8, 24, 36, 37], and more details about the IG dataset are presented in the supplementary materials.

Dataset Comparison. Previous gaze fixation datasets have primarily concentrated on instances (e.g., objects, humans, animals, texts) as targets for recognition, search, and free-viewing, as illustrated in Table 1. In contrast, IG stands as the pioneering interaction-centric gaze fixation dataset. Given the inherent diversity, composability, and context dependency of interactions, interaction-oriented attention is

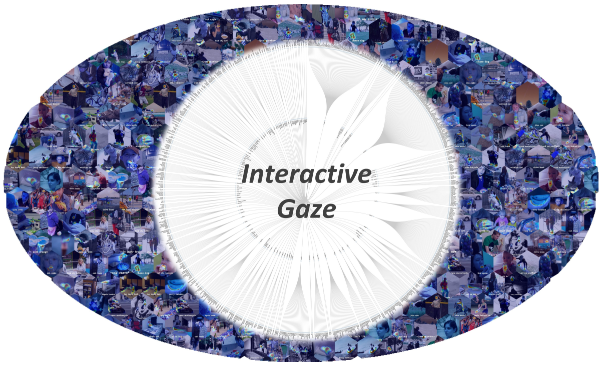

Figure 2. Our proposed IG is the first interaction-centric gaze fixation dataset, comprising 530K fixation points across 740 interaction categories.

Table 1. Comparison of fixation datasets. Previous work all focus on instances of objects, humans, animals, texts, etc.

Dataset	Focus	Scene	Class	Fixation
SALICON [24]	Object	10000	-	4600K
POET [44]	Object	6270	10	178K
People900 [7]	Human	912	1	55K
MCS [58]	Object	2183	2	16K
PET [9]	Animal	4135	6	30K
COCO-Search18 [55]	Object	6202	18	300K
SalECI [23]	Object & Text	972	13	257K
Our (IG)	Action	6299	740	530K

associated with elevated cognitive challenges and empirical knowledge.

Data Selection & Subsetting. To bring the visual attention and HOI detection domains closer, we select image samples from HOI benchmark dataset VCOCO [14] and HICO-det [2] as experimental materials. VCOCO consists of 24 action classes and 214 HOI classes, while HICO-det is larger and more challenging, covering 117 action classes and 600 HOI classes. We select 4,475 samples from the V-COCO training set (13 samples per HOI class on average) to train the attention prediction model. Additionally, we use 1,104 samples from the HICO-det test set and 720 samples from the V-COCO test set for testing the attention prediction model under ZeroIA and fully supervised interaction-oriented attention prediction tasks, respectively.

3. Method

3.1. Problem Definition

Given a sample of HOI pair, the goal of zero-shot interaction-oriented human visual attention prediction (*ZeroIA*) is to generate an attention heatmap that reflects key visual cues for comprehending the interaction between human and object without any training for this interaction, as shown on the left of Figure 3. A complete HOI sample consists of five components: an RGB image $x \in \mathbb{R}^{H\times W\times 3}$, a bounding box of human b^H, a bounding box of object b^O, an object label C_O, and an interaction label C_I. The output of *ZeroIA* is expressed as a spatial attention heatmap $m_{IA} \in \mathbb{R}^{H\times W\times 1}$.

3.2. Interactive Attention

Under goal-oriented cognitive tasks, humans tend to use their existing empirical knowledge and memory to model cognitive goals, and then progressively perceive and understand visual scenes with the goal. Inspired by the human cognitive process, we propose the *I*nteractive *A*ttention (*IA*) to tackle the ZeroIA problem, as shown in Figure 3.

In IA, we first activate the strong knowledge representation capabilities of CLIP [48] through a clever interaction-oriented prompt block (PB). Secondly, we design two tiny and learnable adapters to model the HOI cognitive task. It generates scene-adaptive knowledge prototypes (KPs) for humans, objects, interactions, and robust visual features. Guided by these knowledge, we progressively perceive and understand the HOI scene. Specifically, we first perceive the human and object through a human-object cognitive block (HOCB), then delve deeper into understanding the interaction that occurs between the human and objects through an interaction cognitive block (ICB), and finally decode the interaction-oriented attention map.

Interaction-oriented Prompt. Given an HOI sample, the interaction-oriented prompt block (PB) provides the visual-linguistic large model CLIP with clever inputs applicable to the HOI cognitive task to activate CLIP's powerful representation capabilities. Due to the natural composability of HOI, PB considers humans, objects, and interactions separately, i.e., constructing human-centric, object-centric, and interaction-centric prompts and extracting their textual representations, respectively.

PB takes *"person"* and object textual label C_O as inputs to CLIP text encoder and extracts the human- and object-centric textual representations, T_H, and T_O. The human and object texts are merged into the interaction-centric prompt, i.e., *"a photo of a person [C_I] [C_O]"* since they are closely related to the interaction. Then, the interaction-centric textual representation T_I is extracted. Correspondingly, CLIP's image encoder is leveraged to extract robust visual features $V = [v_1, ..., v_M]$ to take advantage of the semantic alignment between image and linguistics.

Positional & Visual Adapter. The features extracted from CLIP based on PB are somewhat applicable to the cognitive goal, but this is a fixed holistic representation modeling of the goal, which lacks the ability to adapt to each specific HOI sample, especially to diverse spatial configurations and visual scenes. As a result, we design the tiny positional adapter and visual adapter to provide HOI sample-aware knowledge that enables better modeling of HOI cog-

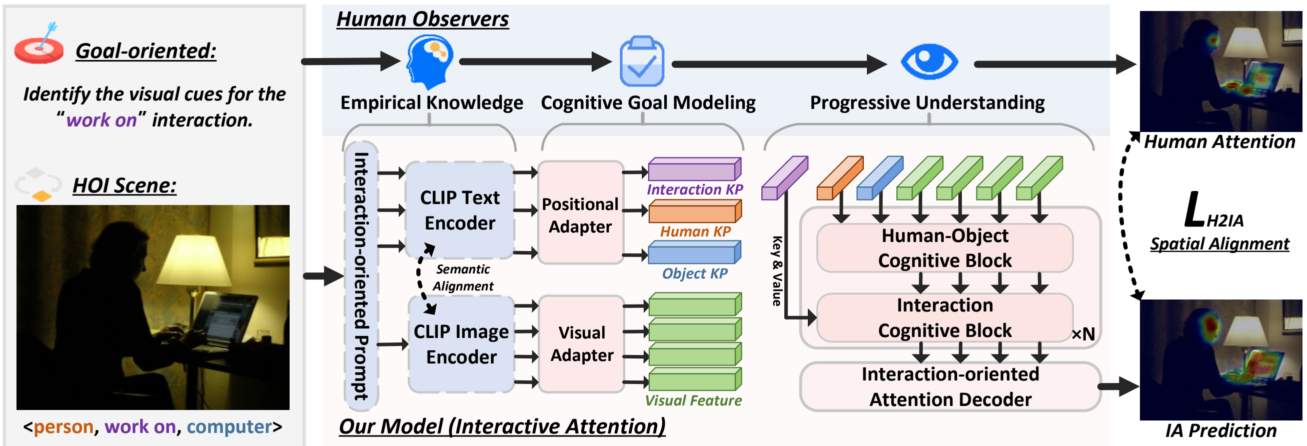

Figure 3. The overall architecture of our *Interactive Attention (IA)*. Inspired by the HOI cognitive process of human observers, *IA* is divided into three phases: empirical knowledge representation, cognitive goal modeling, and progressive understanding. First, a set of interaction-oriented prompts activate and leverage the robust knowledge representation capability of CLIP. Secondly, positional and visual adapters are introduced to acquire scene-adaptive human, object, and interaction Knowledge Prototypes (KPs) along with visual features of the HOI scene. Thirdly, guided by these KPs, *IA* progressively comprehends the scene, starting with an instance-level understanding of humans and objects and deepening insight into their interactions. The decoder generates predicted attention maps, supervised by the real attention maps of human observers using L_{H2IA}.

nitive tasks in each specific sample.

Positional adapter (PA) provides sample-aware spatial information on textual representations:

$$K_H = MLP([(T_H), Fourier(b^H)]), \tag{1}$$

$$K_O = MLP([(T_O), Fourier(b^O)]), \tag{2}$$

$$K_I = MLP([(T_I), Fourier(C_H), Fourier(C_O)]), \tag{3}$$

where K_H, K_O, and K_I represent adaptive knowledge prototypes of human, object, and interaction, $Fourier$ is the Fourier embedding, MLP is a multi-layer perceptron, and $[\cdot]$ is the concatenate operation. Visual adapter (VA) is implemented as two Transformer encoder layers to transform V into refined visual features $V' = [v_1', ..., v_M']$.

Human-Object Cognitive Block. The HOI scene is further learned under the guidance of knowledge prototypes of human and object (K_H and K_O). The self-attention mechanism with gating is introduced to establish cognitive dependencies between knowledge prototypes and visual features, then the human-object-aware visual feature V_{HO} is obtained:

$$V_{HO} = V' + Gate(SelfAttn([K_H, K_O, V'])), \tag{4}$$

where $Gate(\cdot)$ is a token gating operation that considers visual tokens only, $SelfAttn(\cdot)$ denotes a self-attention computation.

Interaction Cognitive Block. Expanding upon the instance-level cognition, including the human and object, the model further delves into comprehending higher-level interactions that occur between the human and object. To achieve this, the cross-attention mechanism with gating is introduced to guide the learning of more subtle interaction cues by interaction knowledge prototype and obtaining the primary HOI-aware visual feature V_{HOI}:

$$V_{HOI} = V_{HO} + Gate(CrossAttn(V_{HO}, K_I)), \tag{5}$$

where we denote $CrossAttn(q, kv)$ as a cross-attention operation. After that, the general self-attention operation is conducted on V_{HOI} to obtain the full HOI-aware visual feature V'_{HOI}:

$$V'_{HOI} = V_{HOI} + SelfAttn(V_{HOI}). \tag{6}$$

Interaction-oriented Attention Decoder. The goal of the decoder is to generate interaction-oriented attention heatmap $m_{IA} \in \mathbb{R}^{H \times W \times 1}$, we first reshape the full interaction-aware visual feature, V'_{HOI}, from a 2D shape of $\frac{HW}{p^2} \times C$ to a standard 3D feature map $\frac{H}{p} \times \frac{W}{p} \times C$, C is the feature channel size, p is the image patch size. Then, we adopt tiny 2-layers of 1×1 convolution with batch norm and ReLU activation to reduce the feature dimension to $\frac{H}{p} \times \frac{W}{p} \times 1$. Last, we adopt a bilinearly upsample operation to achieve the full image size $H \times W$.

Loss Function. We utilize the binary cross-entropy as a loss function for attention alignment, which can be expressed as follows:

$$\mathrm{BCE}(p, y) = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \right), \tag{7}$$

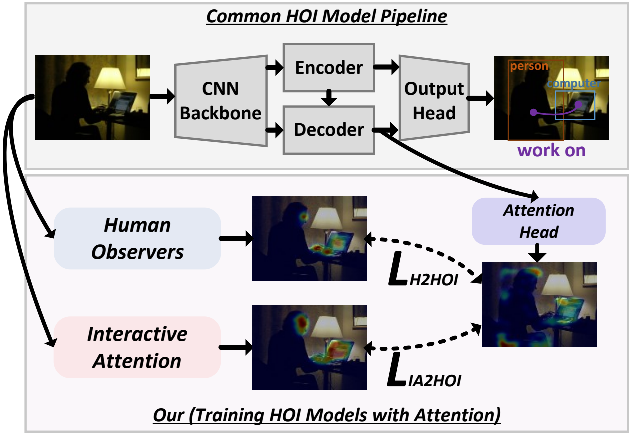

Figure 4. We incorporate aligned attention into the existing HOI model training pipeline, divided into two strategies: supervising by a limited amount of real attention from human observers and supervising by a large number of attention pseudo-labels generated by our proposed *IA*.

$$L_{H2IA} = \text{BCE}(m_H, m_{IA}), \tag{8}$$

where p is the predictive probability, y is the target, N is the number of pixels in m_{IA}, m_H is human observer's real attention heatmap.

3.3. Training HOI Models with Interaction-oriented Attention

Here, we formulate the majority of recent state-of-the-art HOI models in the pipeline of CNN Backbone → Encoder → Decoder → Output Head. We enhance these HOI models by incorporating a supervision of interaction-oriented attention into the training process. This involves two strategies: supervising with a limited amount of real attention samples from human observers and supervising with a large number of pseudo-labeled attention samples from the IA model, as shown in Figure 4.

For these HOI models, we first introduce a simple attention head to extract and reshape the averaged attention maps m_{HOI} from the last layer of cross-attention computation in the decoder. The attention head is intentionally kept simple for a more direct back propagation of guidance to the representation. Subsequently, the corresponding human observer's attention heatmap m_H and IA-generated pseudo-labels m_{IA} are scaled down to match the size of m_{HOI} through adaptive 2D max-pooling. Finally, m_H and m_{IA} are used to supervise m_{HOI} for two kinds of strategies, respectively.

For the strategy of supervising by human observers, the attention alignment loss function L_{H2HOI} and the overall HOI model's loss function L can be expressed as

$$L_{H2HOI} = \text{BCE}(m_{HOI}, m_H), \tag{9}$$

$$L = \lambda_1 L_{raw} + \lambda_2 L_{H2HOI}, \tag{10}$$

where L_{raw} represents the loss function of the original HOI model, which varies depending on the specific model, while λ_1 and λ_2 denote loss scaling factors. Similarly, for the strategy of supervising by *IA*, L_{IA2HOI} and the overall loss function L can be expressed as

$$L_{IA2HOI} = \text{BCE}(m_{HOI}, m_{IA}), \tag{11}$$

$$L = \lambda_1 L_{raw} + \lambda_2 L_{IA2HOI}. \tag{12}$$

4. Experiments

4.1. Implementation Details

Interactive Attention. We apply 2 versions of CLIP (ViT-B/16 and ViT-L/14) to build the IA models, resulting in IA-B and IA-L, respectively. The output dimension of Fourier embedding is 64. During training, we use AdamW as the optimizer with an initial learning rate of 1×10^{-4}. The IA model is trained for 80 epochs with a learning rate reduction pre-20 epochs by a factor of 10. The momentum and the weight decay are set as 0.9 and 1×10^{-4}. The experiments are conducted on four NVIDIA GeForce RTX 3090 GPUs with a batch size of 16.

Training HOI Models with Attention Alignment. We apply our attention-guided training strategy on three typical state-of-the-art HOI methods: MUREN [29], representing the one-stage structure, and UPT [60] and STIP [62], representing the two-stage structure. In MUREN, we align the cross-attention map of the last layer in its interaction branch module with our interaction-oriented attention. In UPT, we first add two transformer decoder layers at the end of its interaction head, conducting cross-attention computation with global visual features, and align the cross-attention map of last layer. In STIP, we align the cross-attention map of the last layer in its structure-aware transformer module with our interaction-oriented attention. Due to variations in the raw losses of different methods, particularly for one-stage structure requiring additional supervision of object detection results, corresponding loss scaling factors differ. λ_1 and λ_2 are set to 1 and 10 in MUREN. λ_1 and λ_2 are set to 1 and 6 in UPT and STIP. Note that the IA model, responsible for generating interaction-oriented attention, is trained without exposure to samples from VCOCO test set and HICO-det test set, preventing any possibility of label leakage.

4.2. Comparison methods & Metrics.

We evaluate the performance of our *IA* model on interaction-oriented attention prediction, comparing it with 10 state-of-the-art methods: ITTI [22], GBVS [17], DeepGaze I [30], DeepGaze IIE [35], UMB [70], ConvNext [40], MLNet [6], SSwin Transformer [23], CLIP-ViT-B/16 [48], CLIP-ViT-L/14 [48]. Among these, ITTI and GBVS

Table 2. Performance comparison on IG dataset. **Bold** and underline show the best and second-best performances.

	ZeroIA Setting				Fully Supervised Setting			
Method	CC↑	KLdiv↓	SIM↑	AUC↑	CC↑	KLdiv↓	SIM↑	AUC↑
ITTI [22]	0.0784	4.3553	0.0142	0.4999	0.0754	3.1815	0.0117	0.5301
GBVS [17]	0.0262	4.5495	0.0234	0.5000	0.1773	2.9498	0.0207	0.5037
DeepGaze I [30]	0.1873	3.0663	0.0080	0.5570	0.1969	3.1869	0.0081	0.6143
DeepGaze IIE [35]	0.2065	3.0255	0.0091	0.5814	0.1721	3.2357	0.0070	0.6049
UMB [70]	0.2121	3.9004	0.5937	0.5326	0.2011	4.2338	0.6147	0.5319
MLNet [6]	0.2937	3.8987	0.4646	0.5972	0.3195	3.6804	0.4153	0.5743
ConvNext [40]	0.2571	4.8729	0.6110	0.5482	0.2731	4.5154	0.5747	0.5524
SSwin Transformer [23]	0.2494	2.6654	0.3309	0.5212	0.2938	3.4309	0.3323	0.5391
CLIP-base [48]	0.1403	3.6118	0.0058	0.3516	0.1564	3.8721	0.0074	0.3385
CLIP-large [48]	0.1681	3.6498	0.0064	0.3183	0.1690	3.8547	0.0066	0.3356
IA CLIP-base	0.4013	**2.7114**	**0.7205**	0.5953	0.4810	2.2132	0.7521	0.6375
IA CLIP-large	**0.4488**	2.8409	0.7152	**0.6068**	**0.5106**	**2.1409**	**0.7796**	**0.6386**

excel in early saliency prediction based on manual features, while the remaining methods are recent learning-based approaches. DeepGaze I and DeepGaze IIE utilize official pre-trained models for predictions, and UMB, ConvNext, MLNet, and SSwin Transformer are all trained with IG datasets. Additionally, we compute the similarity maps of two commonly used raw CLIP versions (CLIP-ViT-B/16 and CLIP-ViT-L/14). These similarity maps compute the similarity distance between image features and text features, serving as the raw CLIP's class attention maps [31].

We apply 4 metrics to measure performance: KL divergence (KLdiv), correlation coefficient (CC), similarity (SIM), and the area under the receiver operating characteristic curve (AUC). KLdiv, CC, and SIM are distribution-based metrics, while AUC is a localization-based metric. Except KLdiv, where lower values are preferable, higher values are desirable for all other metrics.

4.3. Quantitative Comparison.

Table 2 presents the quantitative comparison results in the ZeroIA and the fully supervised settings. We can observe that both versions of our model IA outperform all 10 baselines in both metrics. In particular, for CC metric, IA w/ CLIP-L outperforms the third-best MLNet by 52.8% and 59.8% on the ZeroIA and the fully supervised settings, respectively.

Moreover, we observe that the manual feature-based ITTI and GBVS approaches fail in the interaction-oriented attention prediction task, which demands higher cognitive capabilities. Additionally, when comparing IA with the similarity map of raw CLIP output, our IA achieves significant improvements across all metrics. For example, IA w/ CLIP-B exceeds raw CLIP-B by 186% under CC metric of the ZeroIA setting. This suggests that our model effectively activates and utilizes CLIP's knowledge representation capabilities, adapting them to cognitive tasks on interaction.

4.4. Qualitative Analyses.

As shown in Figures 5 and 6, the visualizations of interaction-oriented attention prediction under ZeroIA and fully supervised settings, respectively. Notably, across both settings, IA exhibits attention distributions that are consistently closer to human observers, outperforming the remaining eight methods.

Table 3. Performance comparison of each modude in IA.

	ZeroIA Setting				Fully Supervised Setting			
Method	CC↑	KLdiv↓	SIM↑	AUC↑	CC↑	KLdiv↓	SIM↑	AUC↑
w/o PA	0.3652	3.2115	0.6733	0.5802	0.3961	2.4676	0.6864	0.6023
w/o VA	0.3838	2.9801	0.6754	0.5832	0.4081	2.4346	0.6893	0.6182
w/o HOCB	0.3498	3.3123	0.6421	0.5743	0.3697	2.5689	0.6484	0.6034
w/o ICB	0.3249	3.6541	0.5845	0.5644	0.3445	2.8533	0.6023	0.5934
IA	0.4013	2.7114	0.7205	0.5953	0.4810	2.2132	0.7521	0.6375

Compared with SOTA models. The previous untrained instance-centric models (columns 6-7) only capture coarse-grained salient persons and objects, but not interaction-related cues. Methods trained on IG prove limited in the complex multi-instance "push" scenario (Figure 5, row 4) under the ZeroIA setting and also in the "talk on" scenario (Figure 6, row 3) with a limited number of training samples.

Compared with raw CLIP. IA accurately captures key visual cues, particularly for unseen categories such as "repair", "kiss", "lick", "push", and "sign" in Figure 5. Noteworthy is the comparison with columns 5-6, where the similarity maps generated by raw CLIP lack interpretability, displaying properties such as opposite visualization and noisy activation, as highlighted in recent work [31, 32]. These observations, along with quantitative results in Table 2, underscore the fact that our work effectively activates CLIP's powerful knowledge representation capabilities and improves interpretability.

4.5. Ablation Study

To examine the impact of each module in IA, we conduct an ablation study by removing modules in Table 3. Specifically, we sequentially exclude the positional adapter (PA), visual adapter (VA), human-object cognitive block (HOCB), and interaction cognitive block (ICB). All ablated models are variations of IA w/ CLIP-B. Results indicate consistent performance degradation across both ZeroIA and fully supervised settings upon removal of any module. Notably, the most significant degradation occurs when the ICB is omitted, emphasizing its crucial role in guiding the learning of interaction-oriented attention by injecting the interaction knowledge. Subsequently, the HOCB, enriched with information about the subjects and objects involved in the interaction, the PA, which incorporates spatial configuration, and the VA, receiving refined visual features, contribute to the model's performance in descending order from high to low.

4.6. Effectiveness of Attention Training Strategy

Aligning human attention enhances HOI model accuracy. We leverage the VCOCO training subset within

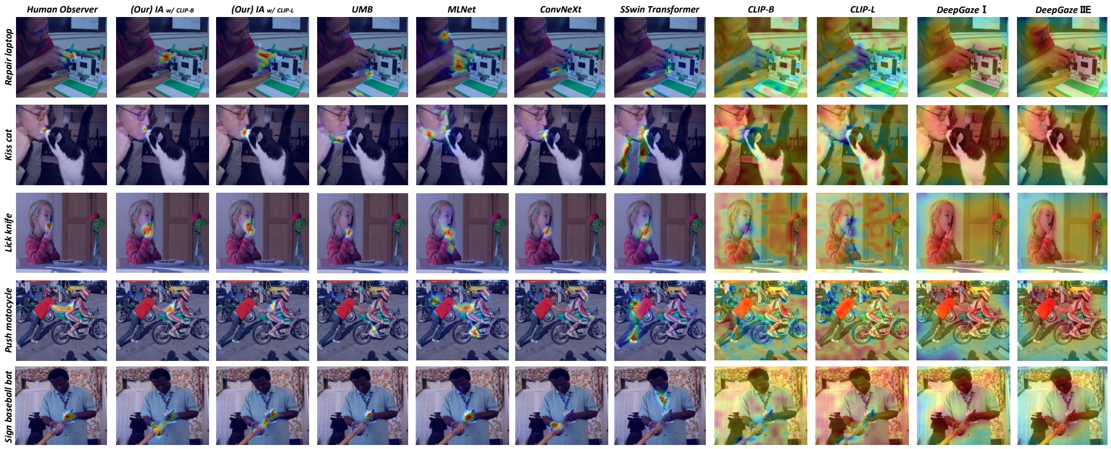

Figure 5. Qualitative comparison of interaction-oriented attention prediction under the ZeroIA setting.

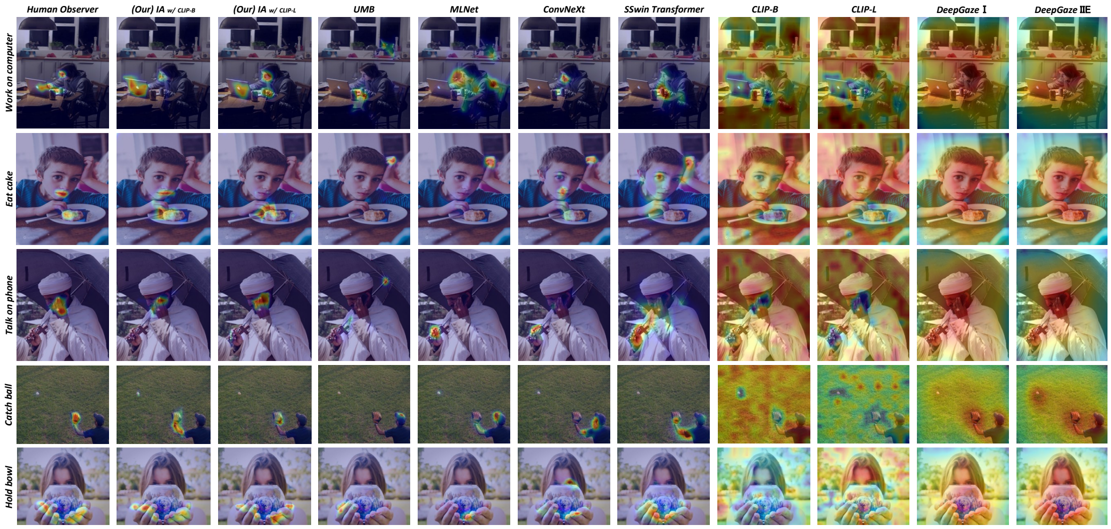

Figure 6. Qualitative comparison of interaction-oriented attention prediction under the fully supervised setting.

the proposed IG, comprising 4,475 samples, approximately 32.4% of the complete VCOCO training set, as the attention ground truth for additional supervision of existing HOI models. As shown in Table 4, MUREN, UPT, and STIP are improved by 0.6mAP, 0.8mAP, and 0.4mAP, respectively, under the Scenario1 setting. Similarly, under the Scenario2 setting, MUREN, UPT, and STIP are enhanced by 0.7mAP, 1.3mAP, and 0.4mAP, respectively. This highlight the significant potential of aligned human attention in enhancing action comprehension, even with limited training data.

Aligning IA-generated attention enhances HOI model accuracy more significantly. We further enhance the additional supervision of HOI models through large-scale interaction-oriented pseudo-labeling, generated by the proposed IA model on the training sets of VCOCO and HICO-det, respectively. In Table 4, on the VCOCO dataset, our training strategies consistently obtain performance improvements for MUREN, UPT, and STIP. In particular, thanks to the precise emulation of human attention by our proposed IA model, the results of all HOI models reveal that training by a large number of pseudo-attention labels is superior to using limited genuine human attention. On HICO-det dataset, our training strategy similarly achieves consistent performance gains in Table 5. Particularly, we observe that our training strategy significantly enhances the performance of the rare set, by 6.0%, 4.0%, and 2.7% on MUREN, UPT, and STIP, respectively. Note that the reported experimental results are all obtained by reproducing these models under the same experimental settings.

Aligning attention makes HOI models more interpretable. We present visualizations of the cross-attention maps of interaction branches in original MUREN and the

Table 4. Comparison of HOI detection performance on VCOCO.

Method	Backbone	Scenario #1	Scenario #2
One-stage methods			
UnionDet [26]	ResNet-50-FPN	47.5	56.2
IP-Net [54]	Hourglass-104	51.0	-
GG-Net [33]	Hourglass-104	54.7	-
HOTR [27]	ResNet-50	55.2	64.4
QPIC [52]	ResNet-101	58.3	60.7
MUREN [29]	ResNet-50	66.9	69.0
MUREN w/ Human	ResNet-50	67.5	69.7
MUREN w/ IA-B	ResNet-50	67.7	70.3
MUREN w/ IA-L	ResNet-50	68.0	70.4
Two-stage methods			
VCL [19]	ResNet-50	48.3	-
SCG [59]	ResNet-50-FPN	54.2	-
UPT [60]	ResNet-50	59.8	65.5
STIP [62]	ResNet-50	66.0	70.7
UPT w/ Human	ResNet-50	60.7	66.4
UPT w/ IA-B	ResNet-50	61.1	66.8
UPT w/ IA-L	ResNet-50	61.4	67.0
STIP w/ Human	ResNet-50	66.4	71.1
STIP w/ IA-B	ResNet-50	67.0	71.3
STIP w/ IA-L	ResNet-50	67.2	71.4

Table 5. Comparison of HOI detection performance on HICO-det.

Method	Backbone	Full	Rare	Non-rare
One-stage methods				
UnionDet [26]	ResNet-50-FPN	17.58	11.72	19.33
IP-Net [54]	Hourglass-104	19.56	12.79	21.58
PPDM [33]	Hourglass-104	21.73	13.78	21.40
GG-Net [63]	Hourglass-104	23.47	16.48	25.60
HOTR [27]	ResNet-50	25.10	17.34	27.42
QPIC [52]	ResNet-101	29.90	23.92	31.69
MSTR [28]	ResNet-50	31.17	25.31	32.92
MUREN [29]	ResNet-50	32.09	27.36	33.50
MUREN w/ IA-B	ResNet-50	32.45	28.90	33.51
MUREN w/ IA-L	ResNet-50	32.48	29.01	33.53
Two-stage methods				
VCL [19]	ResNet-50	23.63	17.21	25.55
ATL [20]	ResNet-50	23.67	17.64	25.47
SCG [59]	ResNet-50-FPN	29.26	24.61	30.65
UPT [60]	ResNet-50	31.93	26.70	33.49
STIP [62]	ResNet-50	32.22	28.15	33.43
UPT w/ IA-B	ResNet-50	32.20	27.59	33.58
UPT w/ IA-L	ResNet-50	32.47	27.76	33.72
STIP w/ IA-B	ResNet-50	32.32	28.47	33.59
STIP w/ IA-L	ResNet-50	32.64	28.90	33.76

MUREN aligned with interaction-oriented attention, respectively, as shown in Figure 7. It is evident that the attention map of 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.

5. Conclusion

This paper aims to address a critical gap in interaction-oriented attention by introducing the challenging yet meaningful ZeroIA problem. To achieve this, we collect the IG dataset, marking the first large-scale interaction-centric gaze fixation dataset, featuring 530,000 fixation points across 740 diverse interaction categories. Subsequently, we present the IA model, ingeniously designed to activate and leverage the knowledge representation capabilities of CLIP, emulating the cognitive processes of human observers to generate high-quality interaction-oriented attention. Extensive experiments demonstrate that the proposed IA model outperforms state-of-the-art attention prediction methods in both zero-shot and fully supervised settings. Furthermore, we propose an attention-guided HOI training strategy, leveraging interaction-oriented attention to guide existing HOI models for more effective learning. Quantitative and qualitative results demonstrate that both genuine human attention and IA-generated attention significantly enhance the performance and interpretability of existing HOI models, making them more capable of visual reasoning like human observers. This highlights the significant potential of visual attention in action comprehension.

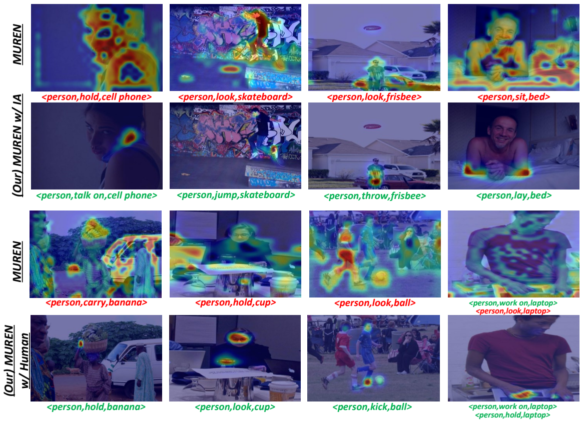

Figure 7. The predicted results and corresponding attention visualizations for MUREN and our (MUREN w/ IA-L, MUREN w/ Human). We mark true positive results in green, and false positive results in red. After aligning interaction-oriented attention, the erroneous prediction results are corrected, and the corresponding attention becomes more converged and more interpretable. More in supplementary materials.

Our endeavor represents the pioneering effort to establish a bi-directional connection between visual attention and HOI detection, starting by directing visual attention through the interaction comprehension task and ending by incorporating this attention to repay existing HOI methods. We believe our work can inspire further research on goal-oriented attention and its application in various action-related tasks like classification, VQA, and captioning. Our work also has potential applications in DNNs' interpretability evaluation, human-AI collaboration, and values alignment research.

Acknowledgements. This work is supported in part by National Natural Science Foundation of China under Grant 62373387, and in part by Natural Science Foundation of Guangdong Province (No. 2023A1515030264).

References

[1] Souradeep Chakraborty, Zijun Wei, Conor Kelton, Seoyoung Ahn, A. Balasubramanian, G. Zelinsky, and D. Samaras. Predicting visual attention in graphic design documents. *IEEE Transactions on Multimedia*, 25:4478–4493, 2023. 1

[2] 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. 2, 3

[3] Shi Chen, Ming Jiang, Jinhui Yang, and Qi Zhao. Air: Attention with reasoning capability. In *ECCV*, pages 91–107. Springer, 2020. 1

[4] Maurizio Corbetta and Gordon L Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews neuroscience*, 3(3):201–215, 2002. 1

[5] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017. 1

[6] Samuel F Dodge and Lina J Karam. Visual saliency prediction using a mixture of deep neural networks. *IEEE TIP*, 27(8):4080–4090, 2018. 5, 6

[7] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 3

[8] 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, 2

[9] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *ICME*, pages 1–6. IEEE, 2015. 3

[10] Chao Gou, Yue Wu, Kang Wang, Kunfeng Wang, Fei-Yue Wang, and Qiang Ji. A joint cascaded framework for simultaneous eye detection and eye state estimation. *Pattern Recognition*, 67:23–31, 2017. 1

[11] Chao Gou, Rui Zhong, and Yuezhao Yu. Mal-net: Multiscale attention link network for accurate eye center detection. *Computer Vision and Image Understanding*, 234:103750, 2023. 1

[12] Chao Gou, Yuchen Zhou, and Dan Li. Driver attention prediction based on convolution and transformers. *The Journal of Supercomputing*, 78(6):8268–8284, 2022. 1

[13] Chao Gou, Yuchen Zhou, Yao Xiao, Xiao Wang, and Hui Yu. Cascade learning for driver facial monitoring. *IEEE TIV*, 8(1):404–412, 2022. 1

[14] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 2, 3

[15] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 1

[16] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *ECCV*, pages 417–434. Springer, 2020. 1

[17] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *NIPS*, 19, 2006. 5, 6

[18] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *ACL*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 3635–3644. Association for Computational Linguistics, 2021. 2

[19] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *ECCV*, pages 584–600. Springer, 2020. 8

[20] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *CVPR*, pages 495–504, 2021. 8

[21] Zhixin Huang, Yuchen Zhou, Jie Zhu, and Chao Gou. Driver scanpath prediction based on inverse reinforcement learning. In *ICASSP*, pages 8306–8310, 2024. 1

[22] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998. 5, 6

[23] Lai Jiang, Yifei Li, Shengxi Li, Mai Xu, Se Lei, Yichen Guo, and Bo Huang. Does text attract attention on e-commerce images: A novel saliency prediction dataset and method. In *CVPR*, pages 2088–2097, June 2022. 3, 5, 6

[24] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 1, 2, 3

[25] Christina Katsini, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. The role of eye gaze in security and privacy applications: Survey and future hci research directions. In *CHI*, pages 1–21, 2020. 1

[26] Bumsoo Kim, Taeho Choi, Jaewoo Kang, and Hyunwoo J Kim. Uniondet: Union-level detector towards real-time human-object interaction detection. In *ECCV*, pages 498–514. Springer, 2020. 8

[27] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*, pages 74–83, 2021. 1, 8

[28] Bumsoo Kim, Jonghwan Mun, Kyoung-Woon On, Minchul Shin, Junhyun Lee, and Eun-Sol Kim. Mstr: Multi-scale transformer for end-to-end human-object interaction detection. In *CVPR*, pages 19578–19587, 2022. 8

[29] Sanghyun Kim, Deunsol Jung, and Minsu Cho. Relational context learning for human-object interaction detection. In *CVPR*, pages 2925–2934, 2023. 2, 5, 8

[30] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze I: boosting saliency prediction with feature maps trained on imagenet. In Yoshua Bengio and Yann LeCun, editors, *ICLRW*, 2015. 5, 6

[31] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 6

[32] Yi Li, Hualiang Wang, Yiqun Duan, Hang Xu, and Xiaomeng Li. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*, 2022. 6

[33] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, pages 482–490, 2020. 8

[34] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *CVPR*, pages 20123–20132, 2022. 2

[35] Akis Linardos, Matthias Kümmerer, Ori Press, and Matthias Bethge. Deepgaze iie: Calibrated prediction in and out-of-domain for state-of-the-art saliency modeling. In *ICCV*, pages 12919–12928, October 2021. 5, 6

[36] 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. 2

[37] Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. In *ICLR*. OpenReview.net, 2019. 1, 2

[38] Xinxin Liu, Yuchen Zhou, and Chao Gou. Learning from interaction-enhanced scene graph for pedestrian collision risk assessment. *IEEE TIV*, 2023. 1

[39] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *CVPR*, pages 3794–3803, 2021. 1

[40] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, pages 11976–11986, 2022. 5, 6

[41] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. Verbs in action: Improving verb understanding in video-language models. In *ICCV*, pages 15579–15591, October 2023. 2

[42] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *CVPR*, pages 1441–1450, 2023. 1

[43] Andrea Palazzi, Davide Abati, S. Calderara, Francesco Solera, and R. Cucchiara. Predicting the driver's focus of attention: The dr(eye)ve project. *IEEE TPAMI*, 41:1720–1733, 2017. 1

[44] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *ECCV*, pages 361–376. Springer, 2014. 3

[45] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, 2022. 2

[46] Warren Woodrich Pettine, Dhruva Venkita Raman, A David Redish, and John D Murray. Human generalization of internal representations through prototype learning with goal-directed attention. *Nature Human Behaviour*, 7(3):442–463, 2023. 1

[47] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual review of neuroscience*, 13(1):25–42, 1990. 1

[48] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2, 3, 5, 6

[49] Cansu Sen, Thomas Hartvigsen, Biao Yin, Xiangnan Kong, and Elke Rundensteiner. Human attention maps for text classification: Do humans and neural networks focus on the same words? In *ACL*, pages 4596–4608, 2020. 1

[50] Baifeng Shi, Trevor Darrell, and Xin Wang. Top-down visual attention from analysis by synthesis. In *CVPR*, pages 2102–2112, 2023. 1

[51] Samy Tafasca, Anshul Gupta, and Jean-Marc Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *ICCV*, pages 20935–20946, 2023. 1

[52] Masato Tamura, Hiroki Ohashi, and Tomoaki Yoshinaga. Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information. In *CVPR*, pages 10410–10419, 2021. 8

[53] Danyang Tu, Wei Sun, Guangtao Zhai, and Wei Shen. Agglomerative transformer for human-object interaction detection. In *ICCV*, pages 21614–21624, 2023. 2

[54] Tiancai Wang, Tong Yang, Martin Danelljan, Fahad Shahbaz Khan, Xiangyu Zhang, and Jian Sun. Learning human-object interaction detection using interaction points. In *CVPR*, pages 4116–4125, 2020. 8

[55] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *CVPR*, pages 193–202, 2020. 1, 3

[56] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *ECCV*, pages 52–68. Springer, 2022. 1

[57] Yushi Yao, Chang Ye, Junfeng He, and Gamaleldin F Elsayed. Teacher-generated spatial-attention labels boost robustness and accuracy of contrastive models. In *CVPR*, pages 23282–23291, 2023. 1

[58] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *CVPRW*, pages 0–0, 2019. 3

[59] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Spatially conditioned graphs for detecting human-object interactions. In *ICCV*, pages 13319–13327, 2021. 8

[60] Frederic Z Zhang, Dylan Campbell, and Stephen Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *CVPR*, pages 20104–20112, 2022. 2, 5, 8

[61] Ruohan Zhang, Akanksha Saran, Bo Liu, Yifeng Zhu, Sihang Guo, Scott Niekum, Dana Ballard, and Mary Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI*, volume 2020, page 4951. NIH Public Access, 2020. 1

[62] Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, and Chang-Wen Chen. Exploring structure-aware transformer over interaction proposals for human-object interac-

tion detection. In *CVPR*, pages 19548–19557, 2022. 2, 5, 8

[63] Xubin Zhong, Xian Qu, Changxing Ding, and Dacheng Tao. Glance and gaze: Inferring action-aware points for one-stage human-object interaction detection. In *CVPR*, pages 13234–13243, 2021. 8

[64] Desen Zhou, Zhichao Liu, Jian Wang, Leshan Wang, Tao Hu, Errui Ding, and Jingdong Wang. Human-object interaction detection via disentangled transformer. In *CVPR*, pages 19568–19577, 2022. 2

[65] Yuchen Zhou, Xinxin Liu, Zipeng Guo, Ming Cai, and Chao Gou. Hktsg: A hierarchical knowledge-guided traffic scene graph representation learning framework for intelligent vehicles. *IEEE TIV*, 2024. 1

[66] Yuchen Zhou, Guang Tan, and Chao Gou. Hierarchical home action understanding with implicit and explicit prior knowledge. In *ICASSP*, pages 4015–4019, 2024. 2

[67] Yuchen Zhou, Guang Tan, Mengtang Li, and Chao Gou. Learning from easy to hard pairs: Multi-step reasoning network for human-object interaction detection. In *ACMMM*, pages 4368–4377, 2023. 2

[68] Yuchen Zhou, Guang Tan, Rui Zhong, Yaokun Li, and Chao Gou. Pit: Progressive interaction transformer for pedestrian crossing intention prediction. *IEEE TITS*, 2023. 1

[69] Yuchen Zhou, Yue Zhang, Zhanwei Zhao, Kaidong Zhang, and Chao Gou. Toward driving scene understanding: A paradigm and benchmark dataset for ego-centric traffic scene graph representation. *IEEE Journal of Radio Frequency Identification*, 6:962–967, 2022. 1

[70] Pengfei Zhu, Mengshi Qi, Xia Li, Weijian Li, and Huadong Ma. Unsupervised self-driving attention prediction via uncertainty mining and knowledge embedding. In *ICCV*, pages 8558–8568, October 2023. 5, 6

[71] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, pages 11825–11834, 2021. 1