Unveiling Causal Attention in Dogs' Eyes with Smart Eyewear

YINGYING ZHAO[*] and NING LI[*], School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
WENTAO PAN, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
YUJIANG WANG[†], Department of Engineering Science, University of Oxford, United Kingdom
MINGZHI DONG[†], School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China
XIANGHUA (SHARON) DING, School of Computer Science, University of Glasgow, United Kingdom
QIN LV, Department of Computer Science, University of Colorado Boulder, United States
ROBERT P. DICK, Department of Electrical Engineering and Computer Science, University of Michigan, United States
DONGSHENG LI, Microsoft Research Asia, China
FAN YANG, School of Microelectronics, Fudan University, China
TUN LU, NING GU, and LI SHANG, School of Computer Science, Fudan University, China and Shanghai Key Laboratory of Data Science, Fudan University, China

Our goals are to better understand dog cognition, and to support others who share this interest. Existing investigation methods predominantly rely on human-manipulated experiments to examine dogs' behavioral responses to visual stimuli such as human gestures. As a result, existing experimental paradigms are usually constrained to in-lab environments and may not reveal the dog's responses to real-world visual scenes. Moreover, visual signals pertaining to dog behavioral responses are empirically derived from observational evidence, which can be prone to subjective bias and may lead to controversies. We

[*] Equal contribution

[†] Corresponding authors

Authors' addresses: Yingying Zhao, yingyingzhao@fudan.edu.cn; Ning Li, 20210240206@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438 and Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Wentao Pan, 21110240007@m.fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438 and Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Yujiang Wang, yujiang.wang@eng.ox.ac.uk, Department of Engineering Science, University of Oxford, Oxford, United Kingdom; Mingzhi Dong, mingzhidong@gmail.com, School of Computer Science, Fudan University, Shanghai, China, 200438 and Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438; Xianghua (Sharon) Ding, xianghua.ding@glasgow.ac.uk, School of Computer Science, University of Glasgow, Glasgow, Lanarkshire, United Kingdom, 200438; Qin Lv, qin.lv@colorado.edu, Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, United States, 80309; Robert P. Dick, dickrp@umich.edu, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, United States, 48109; Dongsheng Li, , Microsoft Research Asia, Shanghai, China, 201203; Fan Yang, yangfan@fudan.edu.cn, School of Microelectronics, Fudan University, Shanghai, China, 201203; Tun Lu, lutun@fudan.edu.cn; Ning Gu, ninggu@fudan.edu.cn; Li Shang, lishang@fudan.edu.cn, School of Computer Science, Fudan University, Shanghai, China, 200438 and Shanghai Key Laboratory of Data Science, Fudan University, Shanghai, China, 200438.

2474-9567/2022/12-ART199 $15.00
https://doi.org/10.1145/3569490

aim to overcome or reduce the existing limitations of dog cognition studies by investigating a challenging issue: identifying the visual signal(s) from dog eye motion that can be utilized to infer causal explanations of its behaviors, namely estimating *causal attention*. To this end, we design a deep learning framework named Causal AtteNtIon NEtwork (CANINE) to unveil the dogs' *causal attention* mechanism, inspired by the recent advance in causality analysis with deep learning. Equipped with CANINE, we developed the first eyewear device to enable inference on the vision-related behavioral causality of canine wearers. We demonstrate the technical feasibility of the proposed CANINE glasses through their application in multiple representative experimental scenarios of dog cognitive study. Various in-field trials are also performed to demonstrate the generality of the CANINE eyewear in real-world scenarios. With the proposed CANINE glasses, we collect the first large-scale dataset, named DogsView, which consists of automatically generated annotations on the canine wearer's *causal attention* across a wide range of representative scenarios. The DogsView dataset is available online to facilitate research.

CCS Concepts: • **Human-centered computing → Mobile devices**.

Additional Key Words and Phrases: Dog Cognition, Canine Vision, Visual Attention, Causal Attention, Dog-Human Interaction, Eyewear Devices

ACM Reference Format:
Yingying Zhao, Ning Li, Wentao Pan, Yujiang Wang, Mingzhi Dong, Xianghua (Sharon) Ding, Qin Lv, Robert P. Dick, Dongsheng Li, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2022. Unveiling Causal Attention in Dogs' Eyes with Smart Eyewear. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 199 (December 2022), 33 pages. https://doi.org/10.1145/3569490

1 INTRODUCTION

Equipped with unmatched social cognition capabilities [28, 51, 75], domestic dogs play a variety of vital roles in modern society: pet dogs relieve stress and improve emotional welfare, guard dogs protect their companions from harm, and seeing-eye dogs assist those with impaired vision; they also protect livestock, remove pests, help their companions hunt, and find lost people [5, 32, 41, 64, 72, 84]. The study of dog cognition is therefore of interest to researchers working in numerous areas [5, 7, 18, 34, 84]. From a scientific standpoint, the investigation of canine cognitive mechanisms can promote a series of dog-related studies, such as Human-Dog Interaction [5, 34, 34, 34, 58, 62, 79], dog's social/non-social cognition [3, 5], and Animal-Computer Interaction [31, 33]. It can also improve our knowledge of human cognition and brains, e.g., the theory of how early humans formulate their social awareness can benefit from the study on canine cognition [26], while our understanding of human cognitive dysfunctions like Alzheimer's disease can also be deepened from the observations of the development of mental deficiencies in dogs [30]. Other academic fields that can benefit from dog cognition studies include the comparative phylogenetics [48] and the ontogenetics [67]. From an application viewpoint, a better grasp of canine cognition can enable improvements in canine training [8], thus enabling dogs to complete their assigned duties better.

Visual perception is arguably the most widely applied cognitive clue in canine cognition research, and visual tasks are the most prevalent paradigm for examining dog behaviors [5]. In those experiments, the experimental subjects, i.e., dogs, are required to perceive the presented visual stimulations. Then, the subject's behaviors in response to those stimuli are recorded and analyzed by human experimenters. This paradigm is commonly applied in canine vision research, including studies of object permanence [13, 22], discrimination learning [2, 54, 55], spatial cognition [11, 20, 21], human-dog interactions [1, 25, 27, 40, 52, 83], etc. A representative example is the classic two-choice paradigm used in Human-Dog Interactions (HDI) [1, 27, 83]. Two potential food sources are provided with a human demonstrator standing between them. One site contains food as a reward, while another does not. The experimenter will then instruct the dog to choose the one with the reward through certain visual human cues, e.g., pointing to the target with arms/legs, gazing at the preferable site, or turning the head. Whether dogs understand these social clues is then measured by their responses, e.g., whether they select the genuine food source.

In the experimental paradigm described above, the entire process is manually controlled, observed, and analyzed by human experimenters. Those human-controlled experiments have considerably expanded our canine cognitive knowledge; however, overly relying on human manipulation can be a double-edged sword. As indicated by Pfungst [63], humans' unintentional cues can influence the behaviors and expectations of social animals, a phenomenon named the Clever Hans effect [63]. This is especially the case in canine studies, as domestic dogs possess highly sensitive mental states and can respond to the subtle gestures made by their owners [53, 81].

There are two main limitations in the existing experimental protocols, e.g., the two-choice paradigm. *Firstly, it is often vulnerable to the subjective bias introduced by human experimenters either consciously or subconsciously.* For instance, if a human demonstrator in a two-choice HDI experiment points to the desired food site with arms or fingers as scheduled, the gaze point may fall onto another site unintentionally. This subconscious behavior can be observed and be prioritized over the pointing gesture by the dog, perhaps leading to an incorrect conclusion, although the correct conclusion would have been reached without the distraction of the demonstrator's gaze. Such unpredictable deviations can occur and accumulate without being realized by experimenters, undermining study reliability. The subjective bias can also result from the overwhelming reliance on human experts to conclude experimental outcomes, as different experimenters can judge the same phenomenon differently. Controversies and debates can emerge as a consequence, e.g., a consensus has not been achieved on the true meaning of the dog's looking-back behavior in the problem-solving task [47], which is interpreted as the indication of asking for help by some researchers but as a result of being attracted by the human's involuntary actions among others.

The predominantly adopted in-lab setting is another significant drawback of the current dog cognitive research. Most experimental paradigms are constrained to in-lab environments, e.g., the aforementioned two-choice protocol. Only human-controlled visual signals are delivered to dogs. This is essentially a compromise of the limited analytical scope of human experts. Dogs in the open world will indeed receive multiple complicated visual signals simultaneously and behave correspondingly. It is nearly impossible for human brains to detect and interpret which visual signal the dog is focusing on and should therefore be associated with the observed behavior. A vast majority of dog cognition studies are designed for laboratories as a result. However, experiments conducted in such a constrained environment may improperly reflect the behavioral reactions of dogs to real-world scenes pertaining to their daily duties. Although several works have already described these problems and have introduced more complex scenarios involving two human instructors [45, 46], the gap between laboratory and real-world conditions remains large.

To relieve the subjective bias and facilitate unconstrained dog cognition research, we propose employing deep learning techniques to understand dog vision from a more objective, accurate, and flexible perspective. Noticeably, the application of deep learning in dog cognition studies, or even in the animal cognition area, is surprisingly rare compared with its impressive success in human visual tasks such as face recognition [16]. To the best of our knowledge, the most relevant works are the applications of eye-tracking techniques to analyze the gazing behavior of dogs [41, 44, 70, 71, 79, 87]. The obtained gaze points are seen as indicators of canine visual attention and the gaze patterns are used to analyze the dog's visual cognition of human faces [4, 71, 72], objects within pictures [70], the implications of oxytocin [73], and so on. Simple as it is from a technology perspective and even without deep networks, the works of this branch have already yielded several interesting discoveries. For example, it is reported by Somppi et al. [70] that canine visual attention generally focuses on more informative regions in pictures, while facial images of conspecifics are preferred over other objects. Despite these findings, the canine eye-tracking apparatus can only detect the gaze point, which can only provide a rough estimate of the dog's observation. Thus, the presented visual stimuli still need to be carefully selected by humans to avoid potential ambiguities. The problem of comprehending real-world canine visual attention in detail remains unsolved.

We take a substantial step towards this beautiful vision by developing a deep-learning-based eyewear device for dogs. This device can analyze the *causal attention* [1] in dog visions, an ambitious task that requires the accurate determination of whether the dog is in a visually attentive state (looking at something attentively) and discovering the visual perceptions that lead to its behavioral response. That is, we would like to identify the *visual attention* that can reveal the causes of a dog's behavioral responses, and the potential benefits are promising. Awareness of *causal attention* allows the canine cognitive experiments to be conducted in unconstrained environments where dogs can behave more naturalistically and realistically. It can also provide an objective evaluation of the specific visual signals that produce particular behavioral responses and, therefore, can significantly eliminate those potential subjective biases in dog cognitive studies.

Revealing such *causal attention*, however, is not a straightforward task. Considering a simple two-choice experiment, i.e., a person points to a ball, and then the dog walks towards the ball as expected. The former can be deemed as the cause of the latter. We first need to explicitly understand what is observed in the dog's eyes, especially in terms of semantic meanings, e.g., ⟨`person` - `pointing to` - `ball`⟩ in this case. Expressing it as a simple triplet graph allows us to discover its causal relationships with the demonstrated behavior. Moreover, there can be multiple such triplet graphs in real-world scenarios, each corresponding to a different visual signal perceived by the dog and serving as a possible cause. Determining *causal attention* signal(s) can be challenging; dogs cannot verbally describe the reasons for their behaviors. We know only the behaviors following visual attention.

This study addresses this problem from an unprecedented perspective, leading to the first eyewear system capable of discovering *causal attention* in dogs' vision. Standing at the heart of this eyewear is a deep neural network named Causal AtteNtIon NEtwork (CANINE). In CANINE, the visual attention of dogs, e.g., human body actions under HDI scenarios, are interpreted as a semantic graph consisting of multiple triplets. Each triplet is in the form of ⟨`subject` - `relationship` - `object`⟩ and stands for a perceived visual signal. Motivated by recent advances in recommendation systems [59], we rely on an effective assumption to estimate *causal attention*. That is, we aim to find the minimum set of the visual attention graph that is most informative to predict the dog's behavioral responses, and we name the resulting set as the *rationale graph* following [59]. The rationale graph is the most compact graph that maximizes the probability of predicting the dog's responsive behaviors.

We design a rationale graph generator that is integrated with CANINE to obtain the rationale graph from a dog's visual attention. The training of CANINE, as a result, has a specific requirement for data annotations, e.g., whether the dog is watching attentively, the type of dog behaviors following an attentive session, and the ground-truth causality, which cannot be found in a public dataset. We first collect a dataset under the HDI scenario to satisfy the training requirement and then manually annotate all the required information. This annotated dataset is used to train the proposed CANINE network. The performance of the resulting CANINE network surpasses that of other baseline methods. To facilitate dog cognition research, we have applied CANINE eyewear to a wide range of representative experimental scenarios, including the two-choice experiment [43], the dog's understanding of misleading signals [19, 49], and a food quantity preference test [64]. Dogs' visual attention and *causal attention* in those sessions are automatically inferred by our CANINE glasses, and we name the resulting dataset DogsView. The DogsView dataset is publicly released to facilitate relevant research on dog cognition, HDI, dog-computer interaction, and more.

To the best of our knowledge, the CANINE eyewear is the first smart glasses device that infers the causes of the canine wearer's behaviors. In the above representative canine cognition experiments, the rationale graph predicted by CANINE eyewear is consistent with the causality estimates of expert human analysts. That

[1] Note that the word "attention" should be distinguished from its usages in the related fields like computer vision [85], which has very distinct semantic meanings. In this work, *causal attention* refers to the process of finding the visual signal(s) that can lead to particular behavioral responses.

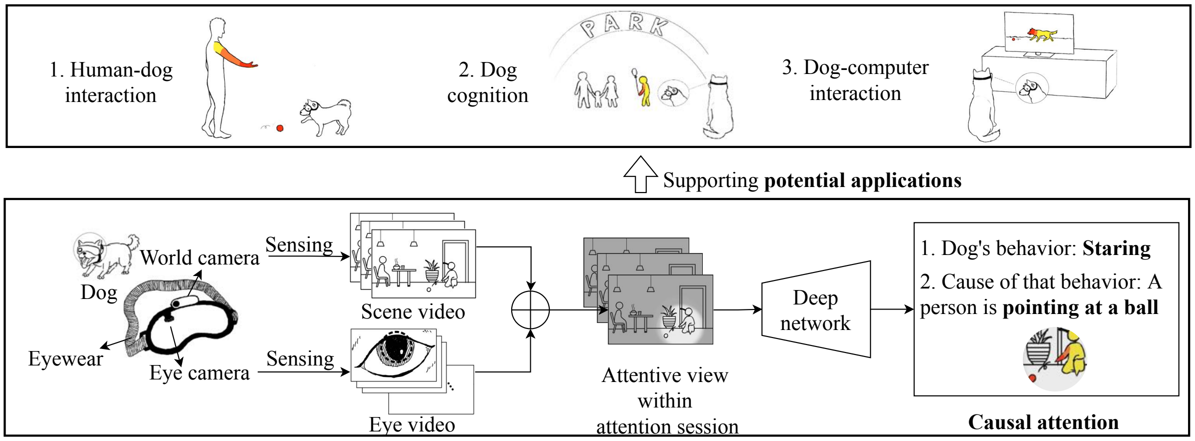

1. Human-dog interaction
2. Dog cognition
PARK
3. Dog-computer interaction
Supporting **potential applications**
Dog
World camera
Eyewear
Eye camera
Sensing
Sensing
Scene video
Eye video
Attentive view within attention session
Deep network
1. Dog's behavior: **Staring**
2. Cause of that behavior: A person is **pointing at a ball**
Causal attention

Fig. 1. The proposed CANINE smart eyewear system.

justifies our assumption of finding *causal attention* with the rationale graph. We have also performed in-field trials to investigate the CANINE glasses' performance under additional real-world scenarios, demonstrating its applicability over unconstrained environments. Figure 1 provides an overview of the CANINE eyewear system.

This study makes the following contributions:

- We propose to analyze the *causal attention* in the canine visual system, which is to discover the causality of the dog's responsive behavior. To the best of our knowledge, this is the first time that such an issue has been explicitly stated and investigated.
- We develop a deep network named CANINE to reveal the *causal attention*, based on the assumption that the rationale graph can appropriately reflect the behavioral causality.
- We develop an eyewear device equipped with the proposed CANINE network to reveal the *causal attention* in dogs' eyes, which opens a gate into dog cognition and behavior studies in the wild. The assumption of the rationale graph has been justified through its extensive applications in various dog cognitive study scenarios.
- We construct the first-ever dataset (dubbed DogsView), featuring dogs' *causal attention* across multiple representative dog cognitive experiments and applications, which is the first to include time-aligned dog eye-area video data and the egocentric scene video data, as well as extensive annotations about dog attention, behavior, and behavioral causality.
- The generality of the proposed CANINE eyewear to unconstrained real-world environments has also been illustrated through in-field trials, exhibiting substantial potential for dog cognitive research.

2 PROBLEM DEFINITION

This paper aims to address the core issue of visual *causal attention* for dogs. However, since most existing techniques and tasks are developed for humans, and the visual systems of humans and dogs are not entirely identical, it is necessary to clarify the key concepts and assumptions in this work to avoid confusion.

Eye Movements As shown in relevant literature [44, 70, 71, 74, 79, 87], the patterns of a dog's eye movements are generally similar to those of a human, which both can be divided into distinct stages of saccades, smooth pursuits, and fixations. Such similarities make the direct application of human eye-tracking systems to dogs a viable option. However, the eye movements of humans and dogs are not identical. A notable difference is that the movements of dogs' eyes, when compared with those of humans, typically exhibit longer fixation periods and shorter saccade durations [74]. This timing difference should be considered when applying human eye-tracking methods to dogs to reflect the biological deviations better.

Visual Attention Although no prior work has explicitly defined the visual attention of dogs, multiple pieces of evidence support the associations between the patterns of a dog's eye movement and its attentive states. Somppi et al. [70] studied whether a dog can focus on the informative region of an image via examining its gazing behaviors. If fixations of eyes are detected, the dog is considered to be attending to the gazing region. On the other hand, smooth pursuit is also a significant indicator of a dog's attention, based on works studying the dog's attention to moving objects [86]. Following those studies, we consider both fixation and smooth pursuit as the proxy for visual attention of dogs, which is generally similar to human attention [12]. However, dogs can neither verbally tell us what they are watching nor perform self-annotations; hence it is challenging to obtain ground truths for dogs' visual attention. Besides, clear definitions of dogs' visual attention remain untouched in prior dog cognitive studies, requiring clarifications from both theoretical and practical perspectives. This work defines canine visual attention as a function of three easy-to-assess conditions: (1) its current eye movement pattern is either smooth pursuits or fixation, (2) the duration of the current gaze pattern has exceeded a threshold, and (3) there is a meaningful *visual stimulus* near the gaze location. The motivations of the first two requirements are to ensure dogs are gazing attentively. They are easily determined by existing smart eyewear. The last criterion excludes cases where dogs are staring at nothing, as it is a clear sign of being mentally unfocused; It also allows us to focus on the central visual field of the dog's vision, which captures at higher resolution than the peripheral visual field. With all three conditions satisfied, it is reasonable to conclude that the dog is attentively looking at something meaningful, i.e., visual attention is established. We can therefore proceed to analyze the attended visual stimuli.

Visual Stimulus The visual systems of dogs and humans are not identical. Physiologic studies have clarified that dogs' eyes are different from humans' in several ways, including color perception [37, 56], sensitivity to light [56, 65], and visual acuity [56, 57], which are not the focus of this work. Instead, we are interested in the implications of visual stimuli and attention processing for dog cognition and behavior studies. Formally, a visual stimulus is defined to be a particular vision, either short-term or long-term, that may lead to a behavioral response. Typical examples include the body actions of the human companion, bouncing balls, flying insects, etc. Dog cognition studies usually categorize those signals into social or non-social cognition [5], depending on whether a human is involved. Without loss of generality, we will mainly employ human body motion in HDI scenarios as the visual stimuli, which fall into the social cognition domain, to illustrate the proposed pipeline of extracting *causal attention*. Other visual stimuli, whether social or non-social, can be addressed likewise.

Semantic Graphs Multiple visual stimuli can be perceived simultaneously. We are interested in determining which signals provide causal explanations for the dog's behaviors. From this perspective, a throughout understanding of the semantic meanings of the visual stimulus is necessary, as it is those semantics that can reveal the underlying causality of the responsive behaviors. For instance, when a human experimenter is pointing in a direction with arms or legs, there can be multiple possible semantic meanings for this gesture. The demonstrator may be pointing to some food or a ball nearby. If the dog perceives the former, it will approach the food, while in the latter case, it will pick up the ball. Merely knowing that a pointing gesture occurs is not enough to determine the cause of the dog's behavior, *it is the semantic meaning of the visual attention that can provide us with a reasonable explanation.* Therefore, we represent each visual stimulus as a triplet semantic graph [38] in the form of ⟨`subject` - `relationship` - `object`⟩, e.g., ⟨`person` - `pointing at` - `food`⟩ and ⟨`person` - `pointing at` - `ball`⟩, which can unambiguously describe behavioral causality. Multiple visual stimuli will be represented as a semantic graph consisting of multiple such triplets.

Causal Attention The concept *causal attention* refers to the perceived visual signal(s) that can be seen as the causal explanation of a behavioral response. With those attentive visual stimuli described as a semantic graph, discovering *causal attention* essentially turns into locating the triplet(s) in the semantic graph of visual attention that can lead to responsive behaviors. In other words, given a scene graph of visual attention, our goal is to find a sub-graph that appropriately describes the behavioral causality. This is an unseen task with two major challenges.

The first one arises from how to determine the ground truths of dogs' *causal attention*, as we cannot rely on the self-annotations from our canine subjects. To address this issue, we have carefully designed the in-lab experiments to identify the causality with high confidence. For example, when a person is pointing to a ball, the dog attends and walks to the indicated ball. Unlike humans, dogs are more mentally straightforward with more predictable behaviors. Thus we can reasonably assume the behavioral causality as ⟨`person` - `pointing at` - `ball`⟩, which needs to be selected out of the graph of visual stimuli.

Obtaining the causal triplet(s) is another difficulty. Few studies have investigated the proper associations of potentially complicated visual signals with behaviors in the real world or determined their causes. Inspired by recent progress in recommendation systems [59], we follow an innovative perspective to address this issue. In particular, we find the *rationale graph* of visual attention, i.e., the minimum set of the semantic graph that can maximize the predictivity for the dog's behavioral responses. The intuition is similar to finding the Markov boundary [60] of the target (behaviors in this case) under assumptions in [60]. The learned rationale graph contains substantial evidence of behavioral causality, and we have shown that this rationale graph indicates the *causal attention* in most dog cognitive study scenarios.

3 SYSTEM DESIGN

3.1 CANINE Network

3.1.1 Overall Pipeline. The general framework of the proposed CANINE network is illustrated in Fig 2. The inputs consist of 1). the dog's eye regions captured by the inward-facing eye camera, and 2). the scene images that are shot with the outward-facing world camera. The first step is to analyze gaze time-series and scene images to find the durations in which the dog is visually attending to something within the scene. For each detected visual attention session, we apply visual masking to each scene image to highlight the regions near the gaze points and to reduce the significance of less relevant parts.

The masked scene images are subsequently fed into a Scene Graph Generator (SGG) [77] to extract the score matrix (a matrix containing posterior probabilities of all semantic triplets) of scene graphs, and then we adopt the BERT model [17] and a temporal pooling method to refine the score matrix and to aggregate the temporal information to a fixed value. The resulting score matrix contains all the possible scene semantics that can lead to the dog's responsive behaviors, and therefore it is defined as the *complete set.* A named *complete predictor* is designed to predict the dog's behavior from the complete set.

Following Pan et al. [59], we train a *rationale generator* to reveal the rationales from the complete set. The output of the rationale generator is essentially a binary mask (or rational mask) of the same size as the complete set, and this mask will be multiplied element-wise with the complete set to obtain a masked set, which is named the rationale set, or the rationale score matrix in particular. This rationale set, as illustrated previously, is intended to be the minimum set that can maximize the predictivity of the dog's behavior. For this purpose, we attach a *rationale predictor* to the rationale score matrix to predict the dog's behavior, and we learn the rationale generator to minimize the prediction gap between the complete predictor and the rationale predictor, following Pan et al. [59].

When produced by a well-trained rationale generator, the rationale set is the most compact set that consists of the maximum amount of semantic information regarding the dog's responsive behavior, and it can be visualized in the form of scene graphs, or the rationale graph, more intuitively. We adopt a Multi-Layer Perceptron (MLP) to summarize the rationale score matrix into a single human action triplet (since we focused on Human-Dog

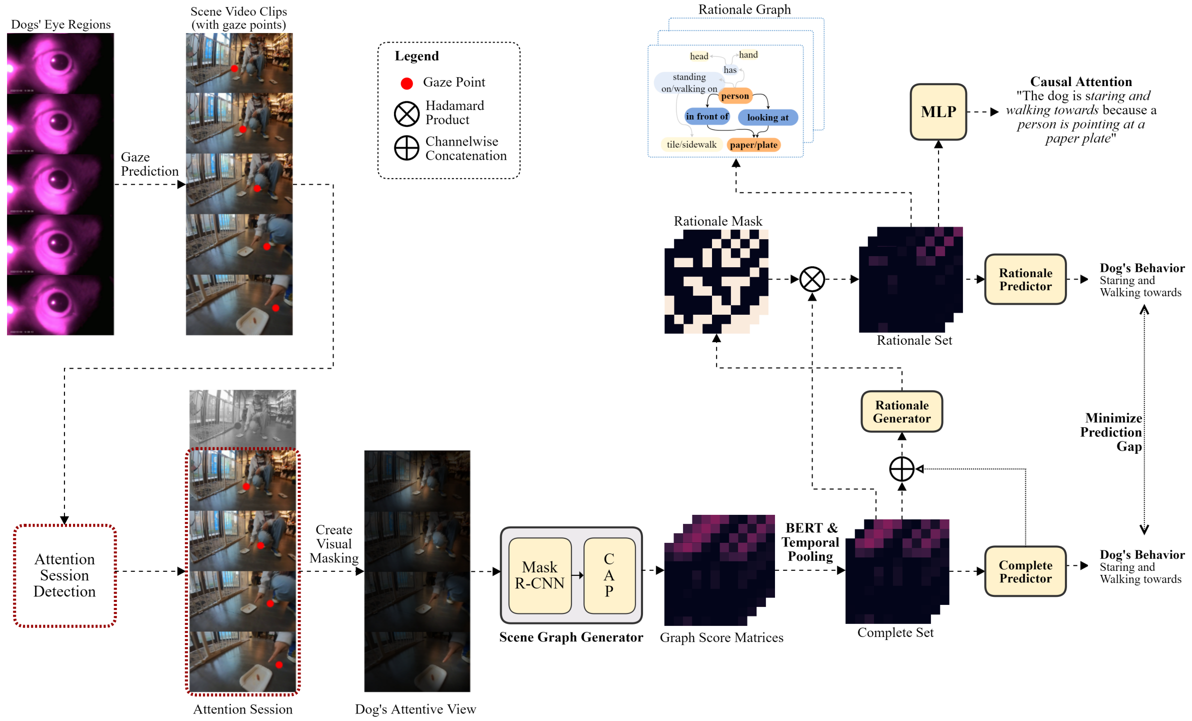


Fig. 2. The framework of CANINE Network.

Interactions, the causality can be assumed as a human action triplet) such that the *causal attention* can be expressed in a short, human-readable sentence.

3.1.2 Attention Session Detection. The first step of the CANINE pipeline is to find the temporal sessions during which the dog is visually attentive. The stage is achieved by analyzing the dog's gaze patterns and the scene images. At a time step t, we consider video frames in time range $[t-N, t+N]$ to be the most relevant ones at this time step and use information from this range to predict the attentive state in step t. Let $\mathbf{E}^t = \{\mathbf{E}_{t-N}, \mathbf{E}_{t-N+1}, ..., \mathbf{E}_{t+N}\}$ be the dog's eye image sequence of this range where $\mathbf{E}_i$ refers to the eye image at step i, and let $\mathbf{I}^t = \{\mathbf{I}_{t-N}, \mathbf{I}_{t-N+1}, ..., \mathbf{I}_{t+N}\}$ denote the scene sequence of the same range where $\mathbf{I}_i$ is the scene image at step i. We first employ Pupil [42] [2] to obtain the gaze position from each eye image. Let $\mathbf{g}_i$ be the gaze position at time step i, and from the eye sequence $\mathbf{E}^t$ we can obtain a sequence of gaze points $\mathbf{g}^t = \{\mathbf{g}_{t-N}, \mathbf{g}_{t-N+1}, ..., \mathbf{g}_{t+N}\}$. We also employ a Scene Graph Generator (SGG) [77] to the scene sequence $\mathbf{I}^t$ to extract a score matrix from each image, i.e., $\mathbf{s}_i = \mathcal{N}_{SGG}(\mathbf{I}_i)$ where $\mathbf{s}_i$ represents the score matrix at step i and $\mathcal{N}_{SGG}$ denotes the SGG network, and denote the score matrices of the sequence as $\mathbf{s}^t = \{\mathbf{s}_{t-N}, \mathbf{s}_{t-N+1}, ..., \mathbf{s}_{t+N}\}$.

With gaze points $\mathbf{g}^t$ and score matrices $\mathbf{s}^t$ obtained, we aim to predict whether dog is visually attentive at time step t. Specifically, we use Temporal Pooling (TP) to score matrices $\mathbf{s}^t$ to summarize the temporal information in the scene sequence, while we also compute the differences between consecutive gaze points to summarize the gaze movement. The temporally reduced score matrix and frame differences of gaze points are concatenated

[2] https://pupil-labs.com

together and then fed into a Multi-Layer Perceptron (MLP) to produce a binary prediction, i.e., whether the dog is visually attentive in time step t.

This binary prediction process is performed in a frame-wise manner, and for each frame we predict whether the dog is visually attentive. A temporal session is seen as an attention session if the dog is visually attentive in most of the frames, which triggers analysis of *causal attention*.

3.1.3 Complete Predictor. After an attention session is detected, the target is to find the dog's behavioral causality in this session. To achieve this, a *complete predictor* is designed to predict the dog's behavior from all possible semantics of the dog's visual attention.

In particular, assume there are a total of K attention sessions during training, let $\mathbf{I}^j = \{\mathbf{I}_0, \mathbf{I}_1, ..., \mathbf{I}_T\}$ be the scene sequence of the j-th attention session of duration $T + 1$ where $j \in [0, K) \cap \mathbb{Z}$, and let $\mathbf{g}^j = \{\mathbf{g}_0, \mathbf{g}_1, ..., \mathbf{g}_T\}$ be the predicted gaze points of the same session. Following MemX [12], we first apply Gaussian Relaxation to convert those gaze points $\mathbf{g}^j$ into heatmaps and apply them as visual masks on the corresponding scene images. The intuition is to focus the model on regions closer to the gaze points, as the locations near the gaze points are areas of potential interest to subjects [12, 61]. This can be denoted as $\mathbf{I}'_i = \mathrm{VM}(\mathbf{I}_i, \mathbf{g}_i)$ where $\mathbf{I}'_i$ is the masked scene image at step i, and VM refers to this Visual Masking processing. The masked scene sequence can be denoted as $\mathbf{I}^{j\prime} = \{\mathbf{I}'_0, \mathbf{I}'_1, ..., \mathbf{I}'_T\}$.

After visual masking, we employ a Scene Graph Generator (SGG) to extract the score matrix of the scene graph from each scene image. We generally follow the implementation of Tang et al. [77], i.e. the SGG consists of a Faster R-CNN [66] and a Casual Analysis Predictor (CAP). Denote the SGG network as $\mathcal{N}_{SGG}$, we have $\mathbf{s}_i = \mathcal{N}_{SGG}(\mathbf{I}'_i)$ where $\mathbf{s}_i$ is the score matrix at step i, and let $\mathbf{s}^j = \{\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_T\}$ be the extracted score matrices for this attention session.

To filter semantically meaningless triplets in the extracted graph score matrices, we apply a BERT [17] model to $\mathbf{s}^j$, i.e., $\mathbf{s}^{j\prime} = \mathcal{N}_{BERT}(\mathbf{s}^j)$, where $\mathbf{s}^{j\prime}$ is the filtered score matrix and $\mathcal{N}_{BERT}$ is the BERT model. Since attention sessions may have different lengths, we use Temporal Pooling (TP) to $\mathbf{s}^{j\prime}$ to reduce the time dimension to be a fixed value (e.g., 10), i.e., $\mathbf{C}^j = \mathrm{TP}(\mathbf{s}^{j\prime})$, where $\mathbf{C}^j$ stands for the temporally reduced graph score matrix and TP refers to the Temporal Pooling operation.

The extracted score matrix $\mathbf{C}^j$ contains all semantically meaningful information that is related to the dog's behavior and is therefore called the *complete set*. We design a *complete predictor*, which is essentially an MLP model consisting of a linear layer and a soft-max layer, to predict the dog's behavior from the complete set. $\mathcal{N}_C$ is the complete predictor. $\hat{y}^j_C = \mathcal{N}_C(\mathbf{C}^j)$, where $\hat{y}^j_C$ is the predicted type of canine behavior based on the complete set. Let y^j be the ground-truth of the dog's behavioral response for this session. A Cross-Entropy loss is defined to train the complete predictor, which can be written as

$$\mathcal{L}_C = \sum_j \mathrm{CE}(y^j, \hat{y}^j_C), \tag{1}$$

where $\mathcal{L}_C$ is the loss of the complete predictor over all attention sessions during training, and CE is the Cross-Entropy loss.

Since the complete set contains all the semantic information, it is heavily redundant, making it difficult for us to retrieve a reasonable behavioral causality. As such, we design a rationale generator & predictor to reveal the pursued *causal attention*.

3.1.4 Rationale Generator & Predictor. As mentioned before, we aim to learn a rationale set, the most compact subset of the complete set, that can maximize the predictivity of the dog's behavior.

Following Pan et al. [59], we first apply a *rationale generator* to generate a binary mask from the complete set $\mathbf{C}^j$ and also from the weights of the linear layer of the complete predictor $\mathcal{N}_C$. The binary mask has the same size as the complete set and indicates whether each element in the complete set should be kept or discarded.

The rationale generator, $\mathcal{N}_G$, is also essentially an MLP model. The weights of the linear layer in the complete predictor $\mathcal{N}_C$ are $\mathbf{W}_C$. We concatenate $\mathbf{W}_C$ with $\mathbf{C}^j$ and then input them to $\mathcal{N}_G$ to obtain a rationale score, i.e., $\mathbf{r}^j = \mathcal{N}_G((\mathbf{C}^j \oplus \mathbf{W}_C))$, where $\oplus$ refers to channel-wise concatenation and $\mathbf{r}^j$ refers to the predicted rationale score. The rationale score matrix $\mathbf{r}^j$ is the same size as the complete set $\mathbf{C}^j$, and each element of $\mathbf{r}^j$ has a normalized value between $[0, 1)$.

Intuitively, a larger rationale score value in $\mathbf{r}^j$ indicates that the element of the same position in $\mathbf{C}^j$ has a higher probability of being the rationale, i.e., being more relevant to *causal attention*. Following Pan et al. [59], we round $\mathbf{r}^j$ into a binary rationale mask $\mathbf{b}^j$, and then multiply it, element-wise, with the complete set $\mathbf{C}^j$ to produce the rationale set, i.e., $\mathbf{R}^j = \mathbf{b}^j \otimes \mathbf{C}^j$, where $\mathbf{R}^j$ stands for the rationale set, $\otimes$ refers to the Hadamard product.

After obtaining the rationale set $\mathbf{R}^j$, we input it into a rationale predictor to predict the dog's behavioral response. Let $\mathcal{N}_R$ be the rationale predictor, which is a MLP model. This can be denoted as $\hat{y}_R^j = \mathcal{N}_R(\mathbf{R}^j)$, where $\hat{y}_R^j$ is the predicted dog behavior based on the rationale set. Similar to the complete predictor's loss, we define a Cross-Entropy loss for the rationale predictor as

$$\mathcal{L}_R = \sum_j \mathrm{CE}(y^j, \hat{y}_R^j), \tag{2}$$

where $\mathcal{L}_R$ is the loss of the rationale predictor over all training sessions. However, $\mathcal{L}_R$ is not the only loss that is used to learn the rationale generator $\mathcal{N}_G$ and the rationale predictor $\mathcal{N}_R$.

3.1.5 Training Objectives. There are two major components to learn in the CANINE network: the complete predictor $\mathcal{N}_C$, and the rationale-related networks, i.e., the generator $\mathcal{N}_G$ and the predictor $\mathcal{N}_R$. Since training the complete predictor $\mathcal{N}_C$ does not reply on $\mathcal{N}_G$ or $\mathcal{N}_R$, we first invoke the loss $\mathcal{L}_C$ in Eq. 1 to train $\mathcal{N}_C$, temporarily ignoring $\mathcal{N}_G$ and $\mathcal{N}_R$.

After learning the complete predictor $\mathcal{N}_C$, we freeze its weights for simplicity and continue to train the rationale generator and predictor. Following Pan et al. [59], we would like to minimize the prediction gap between the complete predictor and rationale predictor since the rationale set should maintain the maximum predictivity from the complete set. This can be written as

$$\mathcal{L}_{R,C} = \mathrm{ReLU}(\mathcal{L}_C - \mathcal{L}_R), \tag{3}$$

where ReLU refers to the ReLU operation. The intuition of Eq. 3 is that the predictivity of the rationale set should be as close to that of the complete set as possible.

We require the rationale set to be a most compact set of the complete set. As such, we add another regularization loss to satisfy this requirement, which can be written as

$$\mathcal{L}_{Reg} = \mathbb{E}(\|\mathbf{R}\|_1) - \eta, \tag{4}$$

where η is a pre-defined gap level, and $\mathbb{E}(\|(\mathbf{R})\|_1)$ refers to the proportion of the non-zero elements in all rationale sets. In other words, we would like the rationale set to be as sparse/compact as possible.

The final loss used to train the rationale generator and rationale predictor can be written as

$$\begin{aligned} \mathcal{L}_G &= \mathcal{L}_R + \alpha\mathcal{L}_{R,C} + \beta\mathcal{L}_{Reg}, && (5)\\ &= \sum_j \mathrm{CE}(y^j, \hat{y}_R^j) + \alpha\,\mathrm{ReLU}(\mathcal{L}_C - \mathcal{L}_R) + \beta(\mathbb{E}(\|\mathbf{R}\|_1) - \eta), && (6) \end{aligned}$$

where $\mathcal{L}_G$ represents the final loss to train $\mathcal{N}_G$ and $\mathcal{N}_R$, α, and β are the weights of different loss terms, respectively.

3.1.6 Obtaining Causal Attention. The learned CANINE network can generate the rationale set $\mathbf{R}^j$ for an attention session; however, it is still one step away from what we would like to achieve: the *causal attention* that can be intuitively understood by humans. Therefore, we further applied an MLP network to $\mathbf{R}^j$, which outputs the type of visual stimulus that can be seen as the cause of the dog's behavioral response. For example, this paper focuses on the scenarios of HDI, and therefore we design the output to be different types of human actions; however, this framework can be easily extended to other dog cognition studies.

With the last piece of the puzzle in place, we are now able to describe the *causal attention* of dogs in clear, human-readable sentences. Since the expected behavior of the dog has already been provided by the complete/rationale predictor, we can assemble it with the behavioral causality and predict the *causal attention* as, e.g., "The dog is *staring and walking towards* because *a person is pointing at a paper plate.*"

The rationale set $\mathbf{R}^j$ is still extremely useful, as it can reveal how the predicted *causal attention* occurs temporally. In this paper, we convert $\mathbf{R}^j$ into a series of temporally consistent scene graphs to intuitively explain how the *causal attention* emerges and changes.

3.2 Hardware Design

The CANINE network is integrated into prototype smart eyewear. The implementation of the prototype hardware generally follows that of EMOShip [89], but we have also made some modifications to suit the use of canines. In this work, we presume that wearing CANINE glasses will not significantly change the behaviors of dogs. To this end, we have employed multiple strategies, including 1) developing a lightweight and comfortable hardware prototype and 2). using a carefully designed training paradigm that allows dogs to acclimate to the eyewear. The hardware of CANINE is retrofitted from commercially available goggles specially designed for dogs with two lightweight cameras: one inward-facing eye camera and one outward-facing world camera.

For the first generation of prototype eyewear, we adopt SHETU SQ11 1080p camera module (1920×1080@30fps) as the eye camera and use an additional IR LED light to light up the iris regions. The world camera is essentially an Insta360 GO 2 (1920×1080@30fps) camera that collects the visual contents adjusted with the canine's FoV. We capture 30 frames per second in this work, as this is adequate for detecting dig eye movement patterns such as fixations, the durations of which are typically longer than 100 ms [61, 74]. A higher frame rate would unnecessarily reduce battery lifespan. The video streams from the two cameras are synchronized by aligning their time stamps, following Pupil [42]. Figure 3 (a) and (b) show the front and lateral views, respectively, of a dog wearing CANINE eyewear.

The smart glasses must be lightweight for the sake of comfort. We developed a newer CANINE glasses prototype in which the mechanical design is overhauled to reduce weight, optimize mechanical reliability, and improve comfort. A key modification is detaching the battery and the control board from the headset and organizing them into a small backpack carried by the dog. This modification reduces the weight of the headset to improve comfort. The resulting headset weighs approximately 72.4 g. The weight of the backpack is 369.8 g. Figure 3 shows a dog wearing the CANINE eyewear.

The proposed system is essentially an offline one. The headset records the video streams with the two cameras, and the recorded data are uploaded to the cloud infrastructure that processes those computation-intensive tasks such as model inferences.

4 DOGSVIEW DATASET CONSTRUCTION

We apply the proposed CANINE system to construct a dog egocentric vision dataset, called DogsView, to support research communities studying dog cognition, HDI, dog-computer interaction, and more. The DogsView dataset covers a wide range of applications and is designed to overcome key obstacles faced by decades-long research on dog cognition, HDI, and dog-computer interaction, namely (1) the potential subjective bias of manual analysis

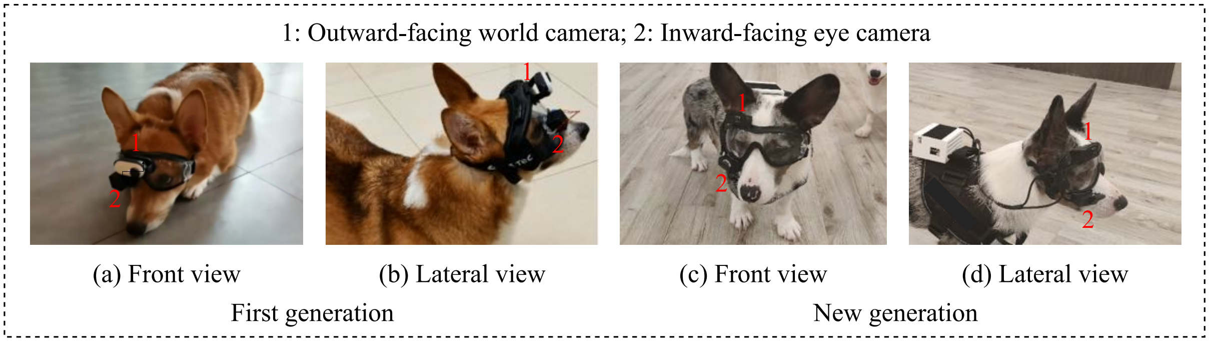


Fig. 3. CANINE eyewear hardware prototype.

results and (2) the limitations of in-lab experiments. Understanding the visual cognitive abilities of dogs is a challenging problem because its scope is broad. This work narrows the problem to automatically capturing visual attention in dogs and discovering the rationale for visual stimuli that lead to their behavioral responses in the open world.

This is the first time that such a dataset has been presented. We have been collecting data over the past two and a half months. The DogsView dataset is available online [15] and we have engaged pilot HDI researchers to support their research using the dataset. As an ongoing effort, we will continue to collect and enrich the DogsView dataset, covering more open-world scenarios and richer HDI signal annotations. It is our hope that the CANINE system and the DogsView dataset will enable interdisciplinary HDI research. All experimental procedures are approved by the ethical committee at Fudan University.

The DogsView dataset is summarized as follows.

- The dataset targets representative applications regarding dog cognition, HDI, and dog-computer interaction. We record dog egocentric scene videos and eye-tracking videos with aligned timelines for representative applications. The egocentric scene videos cover dog visual scenes, which are used for visual semantic understanding by extracting objects (e.g., human subjects) and interactions in dog visual scenes using image and video analysis. Eye-tracking videos capture dog visual attention. Scene and eye-tracking videos are aligned spatially and temporally to reveal dog visual attention over time, identify the corresponding HDI signals, quantify the rationale score of each signal, and then support canine visual causal reasoning.
- The dataset is automatically annotated in terms of frame-by-frame visual attention of dogs, human communication signals (i.e., body actions) and representative scene graphs for each attention session. It contains more than twenty types of human communicative signals in the context of human-dog interactions. These signals contain the seven human communicative signals that have been identified and extensively studied in prior work, such as hand gestures [58] and eye-gaze contact [58]. They also contain 14 new signals that have been mostly ignored due to the limitations of existing studies. This rich set of features may enable researchers to gain insight into dogs' visual cognition by comprehensively and quantitatively measuring how human communicative signals affect dog behaviors, and further reveal why and how dogs continually learn from unseen signals.

4.1 Training Paradigm

We follow the training paradigm of Park et al. [74] to acclimate a dog to wearing glasses and maintain its mental status in appropriate conditions. In each data collection session, we ask an experimenter to bring the dog to the target location. At the beginning of the session, the dog is allowed to act freely, typically barking and running, for around 20 minutes to adapt to the new environment. We then place the CANINE system on the dog and reward the dog for wearing the CANINE eyewear with treats and human praise. It is allowed to acclimate to the device to reduce disturbance from wearing the hardware. After the dog behaves naturally for a few minutes without any stress signals, we turn on the CANINE system and start data collection. The human experimenters observe the behaviors of the dog to ensure it is in an appropriate mental state during the data recording of CANINE. If any unusual signals are spotted, like the dog lying on the ground even after it is rewarded, we immediately stop the recording and take corrective actions (e.g., replacing the tired dog with the one full of energy). Figure 4 (left) shows a dog interacting with a human participant. Figure 4 (middle) and Figure 4 (right) show an example of the captured eye images and scene images with aligned timelines, respectively.

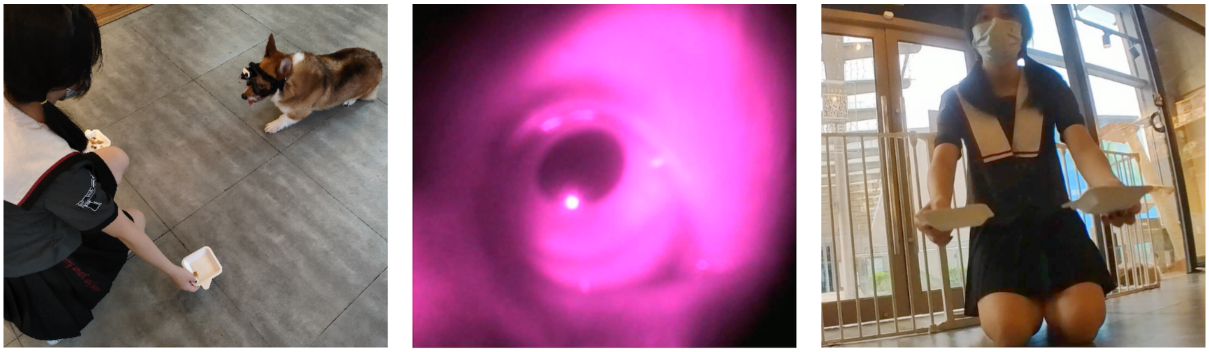


Fig. 4. A photo of data collection procedure (left); Example image frames of the dog's eye image (middle) and the scene image (right) with aligned timelines.

4.2 Applications

The current version of the DogsView dataset covers the following three representative applications related to research on dog cognition, human-dog interaction, and dog-computer interaction, summarized as follows.

4.2.1 Application I: Human-Dog Interaction. Human communicative signals that are easily perceived by dogs and trigger their interactive behavioral responses have long been studied. However, the types of signals identified and extensively studied are limited [5], mostly focusing on hand pointing [58], eye-gaze contact (or eye glances) [79], back-turning [34], and arm extension [34]. In addition, prior research has studied whether dogs can recognize misleading communicative signals from humans [62] and continuously learn new human communicative signals [34]. For example, a previous study found that dogs do not blindly follow misleading human behaviors, such as pointing gesture [62].

Application I is designed with two goals: (1) exploring the visual cognitive processes in dogs perceiving, encoding, and processing misleading communicative signals from humans, and further inferring why dogs can recognize these signals and (2) automatically and quantitatively explaining how dogs learn from new/unknown human communicative signals, and inferring possible causes. We designd Application I as a two-step data collection process. The first step involves manually designed human communicative signals, including actual, misleading, and new/unseen human communicative signals for dogs. In contrast, the second step is exploratory. That is, we take dogs into unconstrained open-world scenarios to collect data on their visual attention, aiming to explore and discover potentially interesting but unknown human interaction cues.

4.2.2 Application II: Dog Cognition. In addition to Application I, which focuses on the social aspect of dog cognition, we also investigate the non-social aspect of dog cognition related to visual stimuli. Application II explores how dogs perceive the physical stimuli [5] of their surroundings, resulting in decisions producing behavioral responses. The importance of non-social aspects of dog cognition has been demonstrated in the past in laboratory settings, but the scope of the previous work is limited in physical stimuli and environmental diversity [5].

In Application II, we instruct the experimenter to take the dog outdoors to experience various daily life environmental scenarios. During this process, the experimenter and the dog interact with the environment freely. During the process, CANINE is used to collect the visual attentive content of dogs in the open world, and automatically analyze dog visual attention as well as the reasons behind it.

4.2.3 Application III: Animal-Computer Interaction. The possible scenarios covered by the field studies described in Application II are important but limited. Meanwhile, as an important subfield of human-computer interaction (HCI), research into animal-computer interaction has investigated how dogs interact with video content, which can potentially cover a wide range of real-world social and non-social scenarios and supplement the studies described in Section II.

We take a dog into a room with an LED screen playing short videos collected from an online platform, such as Tik Tok[3] or YouTube[4]. A person familiar with the dog stays in the same room to make sure the dog is comfortable and relaxed. The dog is free to decide whether to watch the video or not. Thus far, we have collected visual attention data for dogs on 14 online video segments with total length of approximately 20.00 minutes. We further classify these videos into four categories based on their content. Three are socially relevant, such as humans, dogs, and other animals. One is non-socially relevant, containing inanimate objects. To enable a fair comparison, we play an equal number of segments, with similar durations, from each video category, i.e., six video segments and four minutes per category on average.To ensure a diversity of visual stimuli, each category consists of three or more scenarios. For example, we play the *animals* category, including four kinds of animals, i.e., tigers, wolves, and ducks.

4.3 Data Statistics

Table 1 summarizes the DogsView dataset. A total of 213.00-minute timeline-aligned eye-scene videos are constructed, including 63 pairs of eye-scene video clips. The scene video has a spatial resolution of 1,920×1088 pixels, while the eye video has a spatial resolution of 1,920×1080 pixels. Both videos have a sampling rate of 30 fps. Figures 5, 6, and 7 show three examples of time-series scene image frames for the above three applications—HDI, dog cognition, and animal-computer interaction, respectively, in the DogsView dataset.

Table 1. Distribution of the DogsView.

Applications	Representative scenarios	Video minutes	# of videos	# of human action categories	Human actions annotated	Attention annotated	Dog behavior annotated
Human-Dog Interaction (HDI)	Various communicative signals from humans for dogs	124.30	38	21	✓	✓	✓
Dog Cognition	In pet stores, shopping malls, and parks	68.27	11	21	✓	✓	✓
Animal-Computer Interaction	Socially relevant and non-socially relevant	20.43	14	13	✓	✓	✓

[3] https://www.douyin.com

[4] https://www.youtube.com

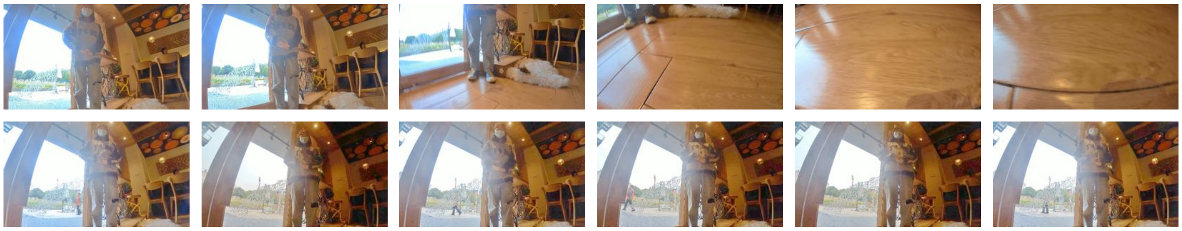


Fig. 5. Time-series image frames of two scenarios in HDI application. Dogs following (top row) and not following (bottom row) misleading communicative signals (“throwing an object”) from an experimenter.

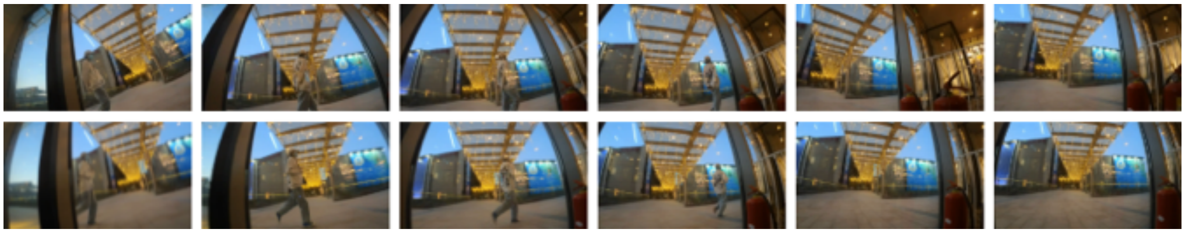


Fig. 6. Time-series image frames of two scenarios in dog cognition application. How dogs visually observe two kinds of passing persons in their private territory (in their home, for example): “walking” (top row) and “running” (bottom row).

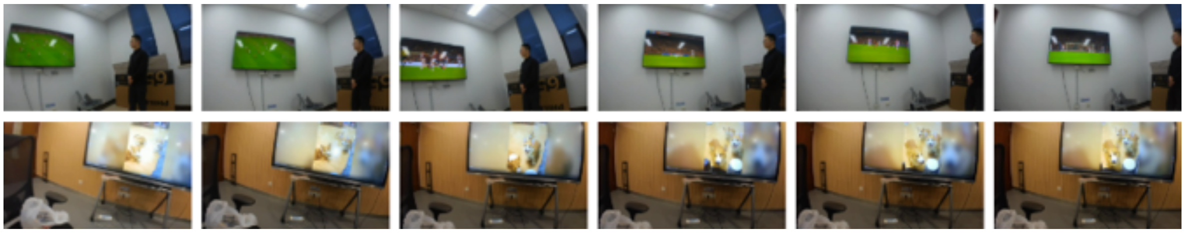


Fig. 7. Time-series image frames of two scenarios in animal-computer interaction application. Dogs prefer to watch videos of the same species (bottom row) compared to other videos (top row).

Figure 8 provides a distribution illustration regarding the ratio of each action type to the number of action signals in the DogsView dataset. There are 21 human action categories, including 7 types of human signals commonly used in previous HDI studies and 14 unintentional signals (marked in blue). The seven types of human signals commonly used in previous HDI studies are: 1) eye-gaze contact [79] or “watch” in this paper, 2) “hand pointing” [58] or “point to” in this paper, 3) “arm extension” [34] (including “carry/hold”, 4) “give/serve (an object) to”, 5) “throw”, 6) hand wave, and 7) hand clap [78]. Those seven signals are mostly related to eye/arm/hand movements, and such actions are carefully selected such that the human experimenters can manually identify and analyze them. Unlike previous works, we devise CANINE eyewear to recognize human actions, which allows for the analysis of more complicated and more subtle visual signals as long as the network has been properly trained. We add a total of 14 new human actions (more signals may be added in the future) that are common in real life but are not well-investigated by previous HDI works. This is also a significant advantage of our system compared to previous HDI works.

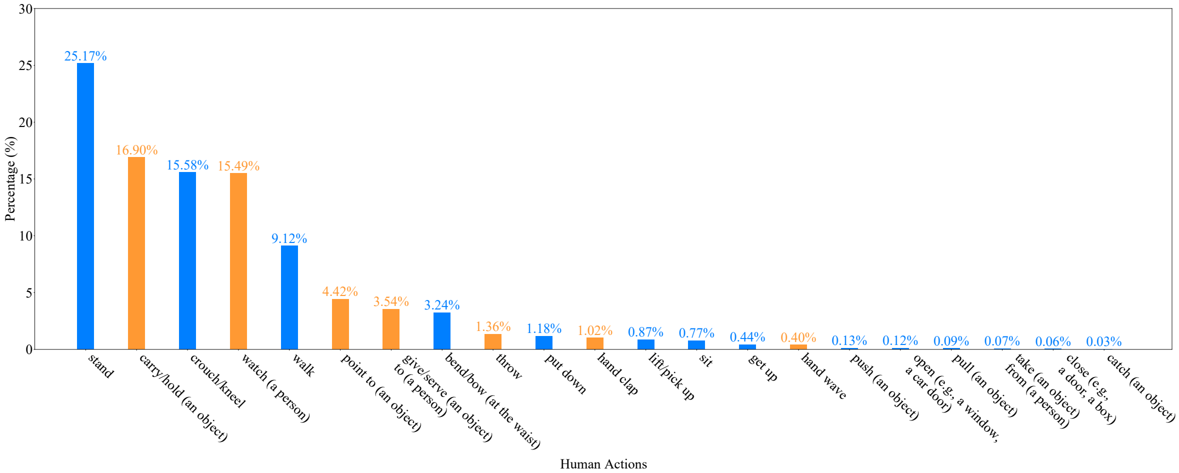


Fig. 8. Illustration of action frequency distribution in the DogsView dataset.

5 EVALUATION

This section reports the performance of CANINE eyewear on three tasks: dog attention recognition, behavior recognition, and causal attention reasoning.

5.1 Experimental Setup

5.1.1 Data Preparation. Before evaluating CANINE, we first need a qualified dataset to train the CANINE network. A qualified training dataset should consist of: (1) time-aligned videos of the dog's eye region and videos of visual scenes reflecting dogs' field-of-view, and (2) annotations of dogs' visual attention and behaviors. As there are no publicly available datasets that can be directly used to train the CANINE network, we need to first collect training data and test data using CANINE eyewear, and label them.

(1) Training and Test Data Collection. We recruit five domestic dogs for data collection to train and test the CANINE system. Specifically, we put the CANINE glasses on each dog and instruct an experimenter to interact with the dog by showing various communication signals, i.e., body actions, that are intended to communicate with dogs. These signals consist of seven intentional human behaviors: "watch", "point to", "carry/hold", "give/serve (an object) to", "throw", "hand wave", and "hand clap". Several involuntary behaviors also occur during data collection. The seven voluntary signals enable us to identify the behavioral causality of dogs with high confidence. In total, we collect 32.46-minute eye/scene video data with aligned timelines.

(2) Labeling Process. Then, we employ two human annotators to manually label dogs' frame-level visual attention, following the same criteria of the dog's visual attentive state defined in Section 2. We also ask human experimenters to observe dogs' behaviors during the experiments and record the category of behaviors. As for the ground truth of behavioral causes, we have carefully designed the experiments such that those causalities can be easily identified with high confidence. For instance, in a two-choice experiment, if the dog rushes to a food site immediately after the pointing action, we can reasonably assume that the action causes its behavior. We also ask the experimenters to determine whether the selected cause is correct based on their experiences. To accelerate the labeling process, we develop a labeling tool using a Python GUI. Figure 9 shows two examples of time-series attentive image frames for a dog with different behaviors.

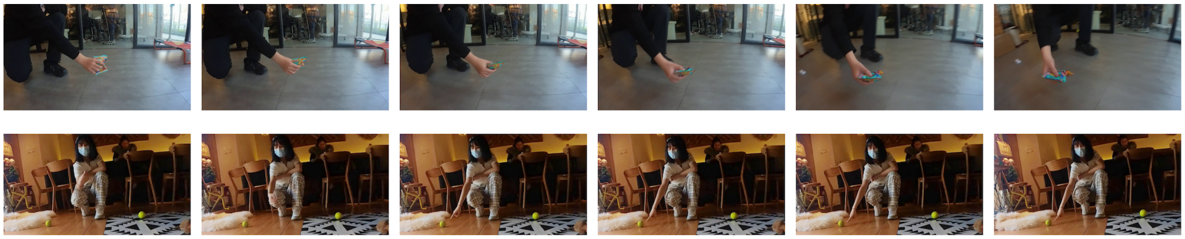


Fig. 9. Examples of two human communicative signals and the corresponding dogs' different behavioral responses. *Top row*: a person is "carrying/holding (an object)"; a dog is "staring and walking towards". *Bottom row*: a person is "pointing to (an object)"; a dog is "staring".

After CANINE is trained, we apply CANINE to examine more scenarios to reveal dogs' visual attention. The data in this stage is used to construct the DogsView dataset stated in Section 4.

5.1.2 Evaluation Metrics. We aim to address an unseen task in this work, i.e., discovering the causal attention in dogs, which is complicated and challenging to evaluate with a single measure. This work mainly examines three aspects of this task: 1). prediction of the dog's attentive state, of which the measures are *Accuracy*, *Precision*, and *Recall*, 2). prediction of the dog's behavioral responses to visual stimulus, which is evaluated with multilabel-based measures, namely *Accuracy*, weighted-averaged *Precision*, and weighted-averaged *Recall*, and 3). finding causes of dog behaviors. The third aspect is difficult to quantify; our analysis is more hands-on in this case. For each experimental session, we visualize the predicted rationale set, a temporally consistent semantic graph, and manually determine whether it has successfully captured the ground truth cause of behavior. The predicted Rationale Score (*RS*), an implicit but quantitative indicator of behavioral causality, is also visualized and analyzed as a side measurement of the prediction quality.

5.1.3 Baselines. Comparisons with prior work are difficult, as the problem of causal attention in dogs has not been considered before. To better understand the effects of CANINE components, we conduct ablation studies.

(1) Visual Attention Recognition. 1) SG-mask method. This method jointly uses eye tracking and scene graph method [77] for attention recognition. In contrast to CANINE, the SG method is set up for an ablation study without using dog visual masks. 2) Eye-tracking method. This method is designed to identify dog visual attention using eye tracking only. The semantic meaning of visual content is not used. It first predicts dog gaze points and then identifies the occurrence of visual attention, i.e., when most gaze points are located in a relatively small region for an adequate duration. This method is similar to the eye-tracking baseline method of Chang et al. [12]. 3) Saliency method, which uses saliency prediction [69] to estimate dog visual attention. The saliency prediction method estimates the regions that attract viewer attention. This task is related to ours, so we adopt it as a baseline methods.

(2) Dog Behavior Recognition. 1) Complete method. Compared with the CANINE network, the complete method is designed to use only the trained complete generator and predictor for dog behavior recognition, in which the complete generator and predictor are disabled. 2) Complete-mask method. Like complete method, the complete-mask method also employs the complete generator and predictor for dog behavior recognition; however, dog visual masks are disenabled. 3) Rationale-mask method. Unlike CANINE, the rationale-mask method does not use dog visual masks.

5.2 Results

5.2.1 Visual Attention Recognition. Table 2 shows the performance comparison of visual attention recognition between CANINE and the baseline methods. CANINE outperforms the baseline methods in terms of *Accuracy*, *Precision*, and *Recall*, indicating the necessity of combining the dog visual mask, dog gaze behavior information, and the semantic meaning of visual stimuli to estimate visual attention.

5.2.2 Dog Behavior Recognition. Table 3 shows the performance comparison of CANINE and the baseline methods for dog behavior recognition. As we can see, CANINE achieves the best performance in terms of *Accuracy*, *Precision*, and *Recall*, demonstrating the effectiveness of using the rationale predictor and dog visual mask.

5.2.3 Causal Attention Reasoning. We summarize the following three representative scenarios from existing work [34, 43, 50, 62] to understand causal attention in dogs.

(1) Whether the dogs follow humans' misleading communicative signals and why. Previous research found that dogs do not always follow misleading human communicative signals [62]. For example, if one holds an object (e.g., a toy ball) when playing with a dog, the dog will stare at one's hand or the object. If one throws the ball, the dog will likely direct its gaze to where it lands. If one only pretends to throw a ball, will the dog's behavior change? What visual stimuli make dogs follow or ignore human "throwing" actions? This scenario aims to answer the above questions by examining how dogs respond to misleading and straight-forward communicative signals from humans.

Table 2. Performance comparison of visual attention recognition.

Method	Accuracy	Precision	Recall
CANINE (Ours)	80.13%	72.31%	70.68%
SG-mask	77.18%	65.16%	68.53%
Saliency	56.21%	40.82%	53.90%
Eye-tracking	45.86%	34.82%	61.56%

Table 3. Performance comparison of dog behavior recognition.

Method	Accuracy	Precision[1]	Recall[2]
CANINE (Ours)	80.27%	79.38%	80.27%
Complete	78.23%	76.89%	78.23%
Complete-mask	66.67%	67.42%	66.67%
Rationale-mask	67.35%	67.84%	67.35%

The experimental procedure follows. We put CANINE on a dog in a room familiar to it. A human experimenter stands in front of the dog, looks at the dog, and throws a small bag. This session includes five trials for each dog. After that, the experimenter repeats the above process in five additional trials. However, during these trials, the experimenter only pretends to throw an object.

We count the proportion of trials in that a dog obeys the two pre-assigned human signals: a straight-forward "throw" behavior and a deceptive "throw", and the results are shown in Figure 10. It can be seen that the five dogs follow the actual "throw" in most trials, while the majority of dogs recognize the experimenter's misleading "throw" and ignore it in most trials.

Figure 11 provides further understanding of how dogs respond to the straight-forward "throw" and the deceptive "throw" by showing the mean Rationale Score (*RS*). *RS* reflects the degree of influence visual stimuli have on dog behavior. We can see that the *RS*s of the deceptive "throw" cases are higher than those of the straight-forward *throw* cases. Intuitively, dogs can recognize the experimenter's deceitful "throw" action, and therefore, they may stare at the human experimenter to gather further information, resulting in higher *RS*. The colored background

[1] refers to weighted-averaged *Precision*

[2] refers to weighted-averaged *Recall*

in Figure 11 indicates the standard deviation of RS in each trial for the five dogs. The variation increases with the number of trials in both the straight-forward "throw" case and the misleading "throw" case. That is in line with our intuition because the dog's physical state differs from trial to trial. In later trials, dogs may become tired. Also, repeated similar communicative signals and interactions with dogs can cause the dog to become fatigued, leading to a big variation in RS.

We also conduct an analysis of variance (ANOVA) to study the effect of different dogs on the RS response in actual "throw" and misleading "throw" cases, respectively. In each case, we make five observations for every dog. The null hypothesis H_0 for the overall F-test is that all five dogs produce the same RS response, on average. The critical value is $F_{4,20} = 2.25$ at $\alpha = 0.1000$. In the actual "throw" case, we obtain $F = 5.35 > 2.25$ ($p = 0.0043$). The result is significant at the 10% significance level. We reject the null hypothesis, concluding that there is strong evidence that the expected values in the five dogs differ. In other words, there is a big individual variation of RS among dogs' responses to the straight-forward "throw". The similar observation can also be found in the misleading "throw" case where $F = 2.58 > 2.25$ ($p = 0.0689$).

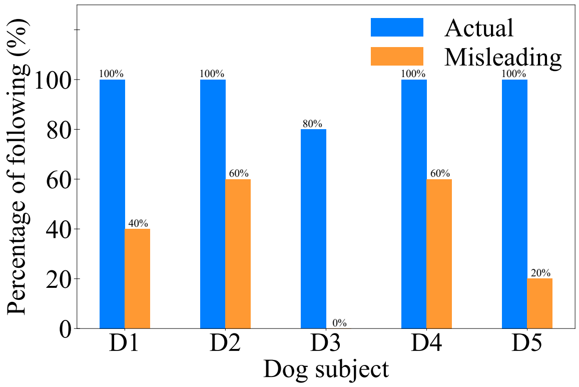


Fig. 10. Percentage of following human actual and misleading signals, respectively.

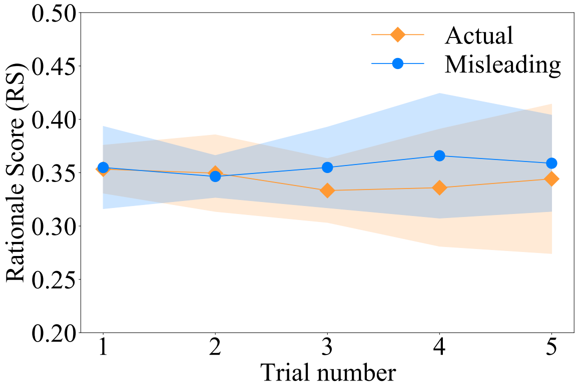


Fig. 11. Variation of mean Rationale Score (RS) when dogs response to human's actual and misleading signals, respectively.

We select two illustrative cases from the above ten trials to show the causal attention reasoning of CANINE, as shown in Figure 12. Figure 12 (top row) shows the following actual action case. From the image frames, we can observe that the dog first stares at the bag in the experimenter's hand, then the dog walks toward the place where the bag falls. The causal attention inferred by CANINE is that the dog is staring and walking towards because of a person's actions: "put down", "watch", "crouch/kneel", and "give/serve". That cause can also be observed from the semantic graphs that contain the triplets explaining the dog's behavior, e.g., ⟨`hand - of - person`⟩, ⟨`hand - holding - bag`⟩. In contrast, Figure 12 (bottom row) shows a case where the dog does not blindly follow the experimenter's deceptive action. We can see that the dog is always staring at the experimenter. The causal attention inferred by CANINE is that the dog is staring because of a person's actions: "stand" "watch", and "walk". The main triplets that explain the dog's behavior are ⟨`hand - of - person`⟩, ⟨`hand - carrying/holding -` ⟩, and ⟨`person - had - leg`⟩.

(2) How the dogs learn from new/unknown human communicative signals. Dogs are known to have the ability to learn and reason continuously. For example, dogs are adept at interpreting humans' communicative signals, such as hand gestures or gaze contact. They can also learn known signals to rapidly understand new ones. This

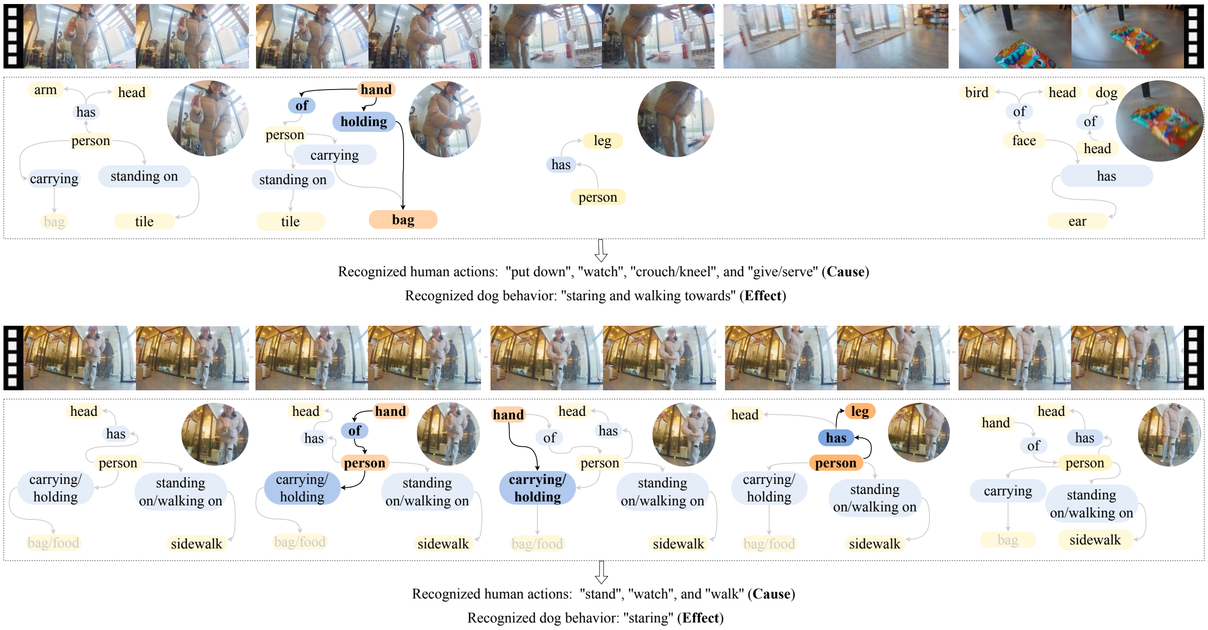

arm
head
has
person
carrying
standing on
bag
tile
hand
of
holding
person
carrying
standing on
tile
bag
leg
has
person
bird
head
dog
of
of
face
head
has
ear
Recognized human actions: "put down", "watch", "crouch/kneel", and "give/serve" (Cause)
Recognized dog behavior: "staring and walking towards" (Effect)
head
has
person
carrying/ holding
standing on/walking on
bag/food
sidewalk
hand
of
leg
has
Recognized human actions: "stand", "watch", and "walk" (Cause)
Recognized dog behavior: "staring" (Effect)

Fig. 12. Two example trials of examining whether dogs follow humans' misleading communicative signals. *Top row*: the dog follows an actual communicative signal "throwing". *Bottom row*: the dog does not follow a deceitful signal.

scenario is designed to examine how dogs perceive known communicative signals and generalize them to new signals. This study compares the known human cue of "hand pointing" with a relatively the under-studied cue of "leg pointing" to understand the dogs' causal attention. More kinds of human cues can be examined in the same way.

The experiment procedure follows. We place two balls on each side of an experimenter. The experimenter stands in front of the dog and looks at the dog. Then, the experimenter uses a finger to points to one of the balls. This process is repeated five times. After that, the experimenter points to a ball with a leg, and the process is also repeated five times.

Figure 13 shows the percentage of trials for each dog following two different human signals. One is "hand pointing", a known signal that dogs are well-known to follow, and the other is "leg pointing", which is an understudied signal that we consider a novel (to the dogs) signal. As expected, the five dogs follow "hand pointing" in most trials (80% or higher), while 3 out of 5 dogs do not follow the "leg pointing" in most trials. We suspect the reason is that the three dogs cannot understand the intention cue from "leg pointing" conveyed by human experimenters.

Figure 14 shows the variation of mean *RS* in these trials. We can see that the *RS*s in "hand pointing" cases are consistently higher than those of "leg pointing" cases. Moreover, the variation of *RS*s for "hand pointing" cases is less than that of "leg pointing" cases. This is because dogs are better at understanding "hand pointing" than "leg pointing", and therefore, dogs will be more attracted by "hand pointing" for gathering communicative information. After ANOVA, $F = 4.09 > 2.25$ ($p = 0.0140$) for "hand pointing" case, and $F = 3.61 > 2.25$ ($p = 0.0225$) for "leg pointing" cases, indicating that there is a large difference in the response degree of the dogs to these two signals.

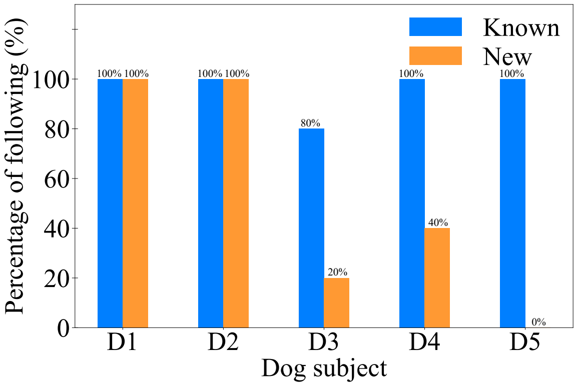


Fig. 13. Percentage of following a well-known "hand pointing" signal and an unknown/new signal "leg pointing", respectively.

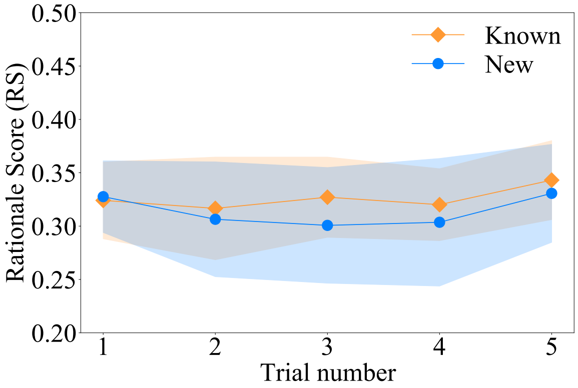


Fig. 14. Variation of mean Rationale Score (*RS*) when dogs response to known and new signals, respectively.

Figure 15 provides two example trials to study how dogs learn a relatively new human communicative signal "leg pointing" from a well-known one "hand pointing". Figure 15 (top row) shows that dog attention focuses on parts of the experimenter's body, e.g., hand, finger, and leg. The causal attention inferred by CANINE is that the dog's behavior is "staring and walking towards" because of humans' actions: "crouch/kneel" and "point to". The semantic graphs provides further insights. For example, in the middle of the attention session, the main triplets reflecting the causes of the dog's behavior are ⟨person - has - hand⟩, ⟨hand - has - finger⟩, ⟨person - in front of - spherical object⟩, and ⟨spherical object - under - hand⟩. Figure 15 (bottom row) provides an example of a "leg pointing" action. The dog first focuses on the experimenter's leg. Then, the dog notices that the leg is close to a spherical object, and the dog walks toward it. Therefore, CANINE concludes that the dog's behavior is also "staring and walking towards". However, it states that the human behavior that causes the dog's behavior is "crouch/kneel".

(3) Do the dogs make counterproductive choices and why. Previous work has shown that dogs make counterproductive choices when they notice humans' ostensive signals [43, 50]. For example, in Prato-Previde et al.'s food quantity preference test [64], if humans show an explicit preference for dogs to choose smaller amounts of food, dogs will ignore their nature of choosing larger quantities and follow humans' preference. We design a similar food quantity preference task to automatically reveal that dogs make counterproductive choices in response to overt cues from humans in an open-world environment.

We prepare two plates of food in a dog's daily room, one with more food and one with less. We alternately show the two plates of food to the dog and place the two plates on two sides of an experimenter. An assistant holds and pets the dog, preventing the dog from eating the food immediately. The experimenter looks at the dog and points to the plate with the smaller amount of food. The assistant then lets the dog go, allowing the dog to choose a plate. This process is repeated ten times. The dog is permitted to rest for a few minutes between trials.

Figure 16 shows the percent of trials in which every dog makes counterproductive choices in the food quantity preference test. It can be seen that all dogs make counterproductive choices in most trials, which demonstrates that, in most cases, dogs ignore their nature of choosing larger quantities and follow human preference. As we can see in Figure 17, mean *RS* is stable, as the number of trials increases despite fluctuation and has less variation than those in the above two scenarios, which demonstrates that there is a relatively low difference in the response degree of dog subjects ($p = 0.1004$).

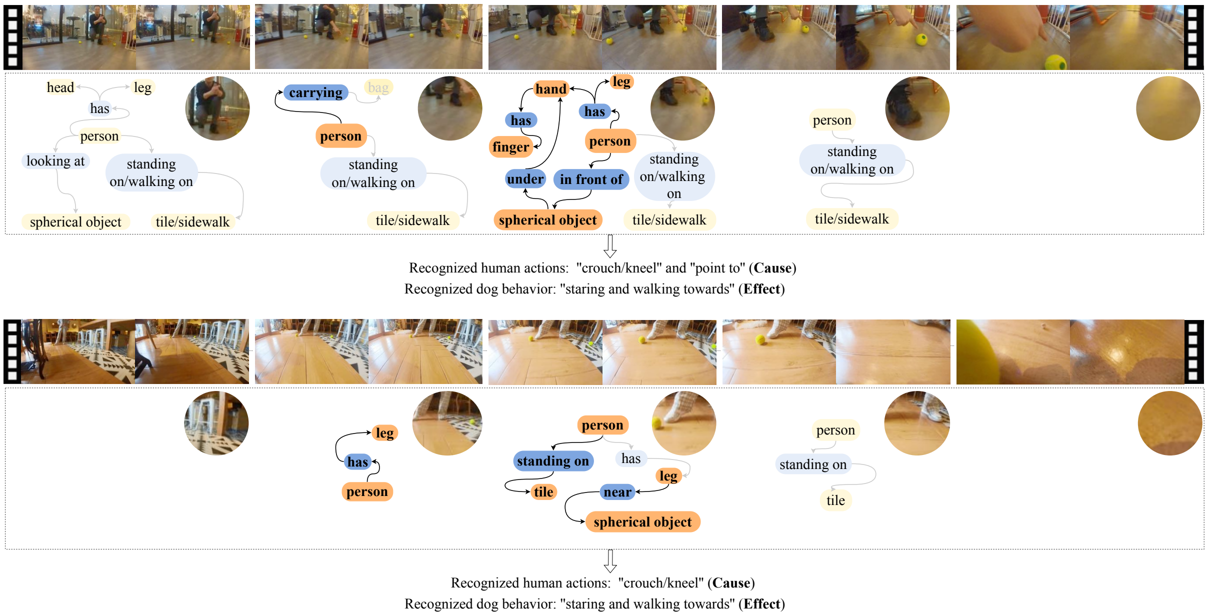


Fig. 15. Examples of the semantic graphs when the dog follows a well-studied known action “hand pointing” (*Top row*) and a relatively new action “leg pointing” (*Bottom row*), respectively.

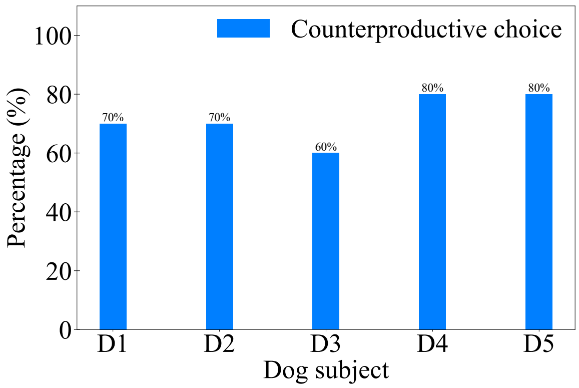


Fig. 16. choice for five dogs in the food quantity preference test.

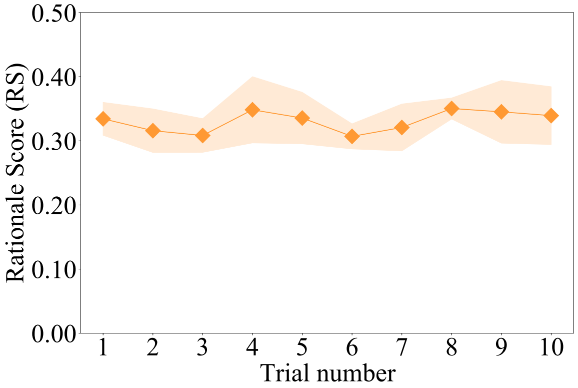
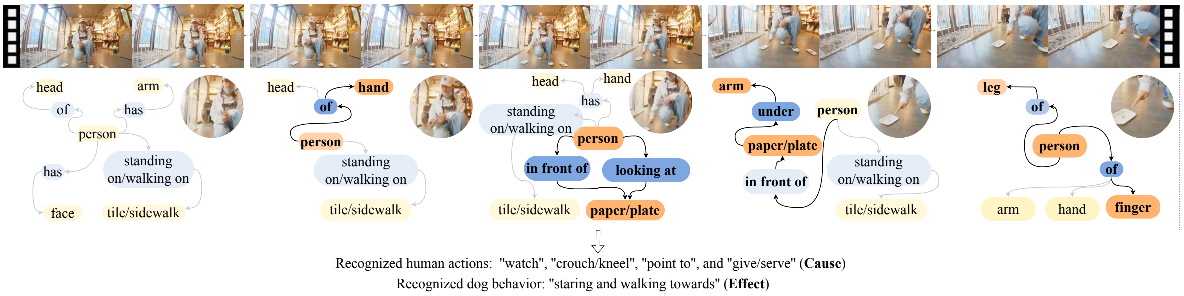


Fig. 17. Variation of mean Rationale Score (*RS*) for five dogs.

Figure 18 shows an example trial of the food quantity preference test. In this trial, the dog looks and walks toward the smaller amount of food, following the experimenter’s explicit body actions, including “watch”, “crouch/kneel”, “point to”, and “give/serve”. We can see that, the main time-series triplets that reflect the cause of the dog’s behavior in the scene graphs are ⟨person - of - hand⟩, ⟨person - in front of - paper/plate⟩, ⟨person - looking at - paper/plate⟩, ⟨paper/plate - under - arm⟩, and ⟨person - of - finger⟩.



Recognized human actions: "watch", "crouch/kneel", "point to", and "give/serve" (**Cause**)
Recognized dog behavior: "staring and walking towards" (**Effect**)

Fig. 18. Examples of semantic graphs generated by CANINE in a trial when the dog makes counterproductive choice.

6 FIELD TRIALS

This section presents the results of field trials to explore the potential use of the proposed CANINE system in open-world scenarios. The CANINE system enables the automatic capture of dog visual attention and provides quantitative causal analysis. CANINE can potentially be used to support various research topics, including human-dog interaction [23, 36], dog cognition [5], dog-computer interaction [31], and more. Based on multi-round interviews with pet owners and researchers, we selected three representative scenarios corresponding to the aforementioned research topics for our field trials:

- dogs in safeguarding roles [68];
- comparing visual attention between dogs and humans [84]; and
- dogs watching videos [33].

Next, we detail the procedures and results of the three field trials and illustrate how the CANINE system can be used to support various research related to visual cognition in dogs.

6.1 Scenario I: Dogs in Safeguarding Roles

Dogs play many roles in our society, and safeguarding has always been one of the most important. Therefore, we chose dog safeguarding as the first scenario. According to related existing research [29, 35], domestic dogs naturally assert and protect their private territory. Haug et al. [29] indicate that domestic dogs usually have territorial behavior in their home and yard. They are generally wary of anyone entering their territory and respond differently depending on factors such as whether they are familiar with the intruder. In this context, studying how and why dogs direct their visual cognitive focuses to different objects can help better understand dog visual cognition behind the safeguarding practices.

6.1.1 Procedure. This study aims to capture how and why dogs allocate visual cognitive resources differently to different people passing by or traversing their private territories. To this end, the experiment consists of three steps. (1) We place the dog indoors, facing a floor-to-ceiling glass window and a glass door to ensure the dog can see people passing by or entering the door. (2) We put the CANINE system on the dog, and let the dog face the window and the door. (3) We summarize a 2 ("familiar person" or "unfamiliar person" with dogs) × 6 ("close to the door, running" or "close to the door, walking" or "far away from the door, running" or "far away from the door, walking" or "entering the door, walking" or "entering the door, running") confounding variable matrix and use all combinations of these variables to examine the various situations a dog might face.

6.1.2 Results. We infer how dogs visually notice people passing by or entering their private territories by computing *RS* for different human actions, including "passing by", "running", and "entering". Table 4 shows the results. The key observation is that dogs show the highest *RS* for familiar participants near the door, and the lowest *RS* for unfamiliar participants running far from the door.

Table 4. Rationale score (*RS*) of different actions for a familiar participant and an unfamiliar one.

	Passing-by				Entering	
	Running		Walking			
	Far from	Close to	Far from	Close to	Running	Walking
Familiar participant	0.28	0.34	0.34	**0.38**	0.33	0.33
Unfamiliar participant	**0.17**	0.36	0.34	0.37	0.33	0.37

The following case further helps to intuitively explain the reasons behind the observation. Take for example the "familiar participant, close to the door" and "unfamiliar participant, far away from the door". Figure 19 shows the time-series of image frames for a dog spotting a familiar passing person who is walking near a door and an unfamiliar person who is running far from the door. As can be seen from Figure 19, the dog has been staring at the familiar person walking near the door. In contrast, the dog does not stare an unfamiliar person when they walk away. A plausible explanation is that the dog does not recognize the unfamiliar person and therefore does not continue to visually inspect them when they are far away from the door.

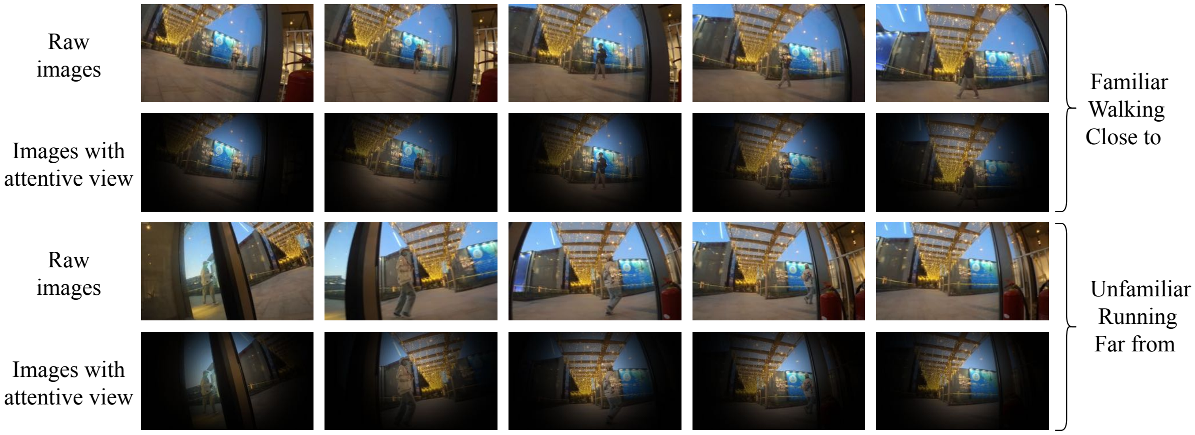

Raw images
Images with attentive view
Familiar Walking Close to
Raw images
Images with attentive view
Unfamiliar Running Far from

Fig. 19. Time-series image frames for a dog when it faces a familiar person who is walking close to a door and an unfamiliar person who is running far from a door.

As we have seen, the proposed system can help support research on dogs in safeguard roles by providing intuitive and quantitative insights into their reasoning based on casual attention.

6.2 Scenario II: Comparing Visual Attention between Dogs and Humans

Comparing visual cognition between dogs and humans has long been an attractive research topic [82]. As shown in previous work [74], dogs have similar yet simpler visual cognitive abilities. However, our understanding of their similarities and differences is still very limited [7]. With the proposed CANINE system, we hope to broaden and deepen our understanding by automatically capturing and quantitatively comparing dog visual attention data in a variety of real-world situations.

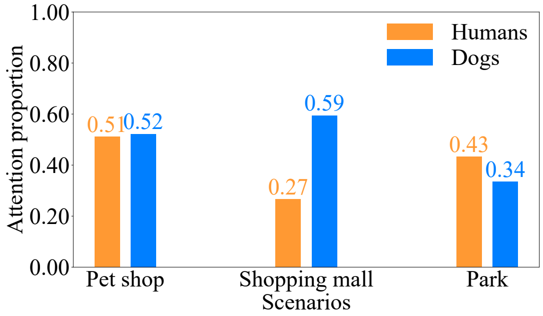


Fig. 20. Comparison of visual attention proportion for dogs and humans in three places.

Section 2 defines dog visual attention, here we define human visual attention in a similar way to yield comparable results. Specifically, we define a human visual attention event to satisfy the following three conditions: (1) the person's eye movement phase is a fixation or smooth pursuit; (2) there is a fixation target (or informative region [70]) that the person is looking at or a moving target that the person's gaze is steadily following; and (3) the fixation or smooth pursuit phase continues for a period of time. To capture human visual attention events, we adopt MemX [12] and slightly adjust its software to meet our requirements.

6.2.1 Procedure. To illustrate how this system might help explore causal attention in dogs compared to humans, we recruit five human participants into the study to experience a variety of everyday environmental scenarios along with dogs. Each human participant is randomly teamed up with a dog that has participated in previous laboratory experiments. We select the following three scenarios to explore how the CANINE system can help reveal their causal attention differences: (1) shopping malls, where dogs may be more attracted to the variety of commodities, e.g., persons or stores, than humans; (2) parks, where humans may find something more appealing than dogs; and (3) pet stores, where it is hard to predict how they will be drawn to different things. We equip the dog and human participants with smart eyewear, and ask each human participant to experience the three environments described while accompanying a leashed dog. In this way, we ensure both the dog and participant experience a similar visual environment and frequently have overlapping fields of view.

6.2.2 Results. Figure 20 shows the visual attention proportion of the attentive image frames to the total frames for dogs and humans participating in this study. Dogs have significantly higher attention proportions than humans (0.59 vs. 0.27) at the shopping mall, but the opposite is true at the park (0.34 vs. 0.43). Interestingly, when they are in the pet store, the number of attentive events are similar (0.52 vs. 0.51).

Their attentive content is significantly different, as shown in Figure 21. In the park, the dog is attracted to persons with more body movements or other inanimate elements, and it ignores the intentional communicative signals from other pedestrians, such as eye contact or waving to the dog. We guess dogs can deliberately ignore human interactive signals and selectively shift attention to their attentive ones.

6.3 Scenario III: Dogs Watching Videos

Animal-computer interaction (ACI) is emerging as a subfield of HCI [33]. Dog-video interaction is a popular ACI research topic, e.g., studying dogs' visual habits while watching videos [33] and understanding how dogs interact with video screens [31]. The CANINE system may facilitate investigation and the design of video interactive technologies for dogs, thereby promoting the dogs' welfare.

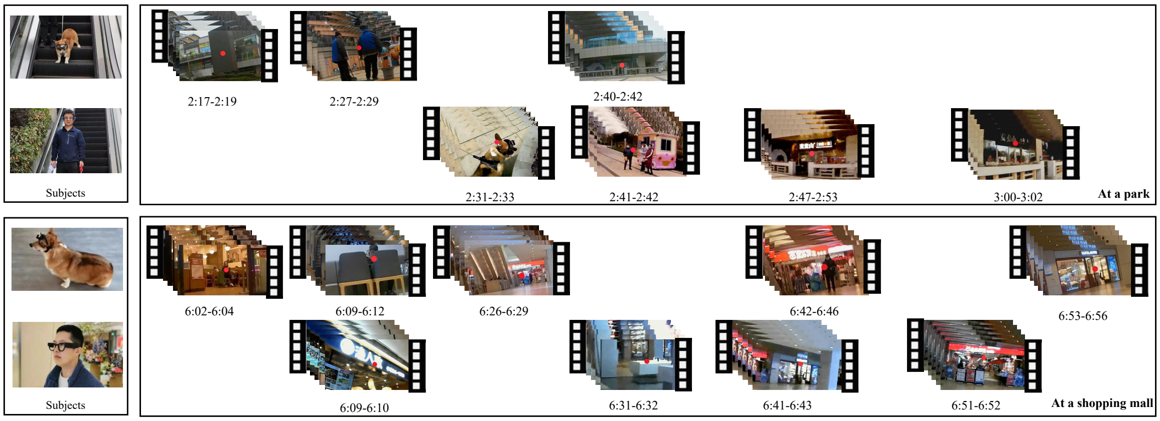


Fig. 21. Visual attention differs in dogs vs. humans at different places.

6.3.1 Procedure. We intend to explore how the CANINE system can automatically capture and measure visual attention when dogs watch videos. We take a dog into a room with an LED display playing short videos from the websites such as Tik Tok[5] and YouTube[6]. A person familiar with the dog is in the same room, making sure the dog is comfortable and relaxed. The dog can freely choose whether to watch the video or not. Figure 22 (left) shows a picture of the dog in this trial.

6.3.2 Results. In this study, we play a collection of short videos with a total length of 20.42 minutes. We found that that the dog only spends 4 minutes watching videos. This result is consistent with our expectations and also the findings reported in related studies [33]. For example, an existing study points out that human-preferred video content may not attract the attention of dogs [32]. Furthermore, the CANINE system allows us to easily identify and quantify how the different video content attracts dog visual attention and potentially explain why. The collected video content is divided into four categories, socially relevant such as *Humans*, *Dogs*, and *Other Animals*, and non-socially relevant such as *Inanimate Elements*. Figure 22 shows the percentage of frames that grab dogs' visual attention as a percentage of the total video frames. We can see that the video categories that attracts dog attention most frequently are *Dogs* first, then *Humans*, followed by *Other Animals* and *Inanimate Elements*. As suggested by relevant studies [5, 6, 18, 70], dogs have inherent social characteristics [5, 6], and while finding their own species more attractive [70], their attention is also attracted by humans due to their long-term co-living with humans [18].

7 DISCUSSION AND LIMITATIONS

This study provides a new way to understand cognitive attention in dogs by using smart eyewear to automatically capture the visual attention of dogs in the open world, and further quantify attention events as well as infer related causes.

[5] https://www.douyin.com
[6] https://www.youtube.com

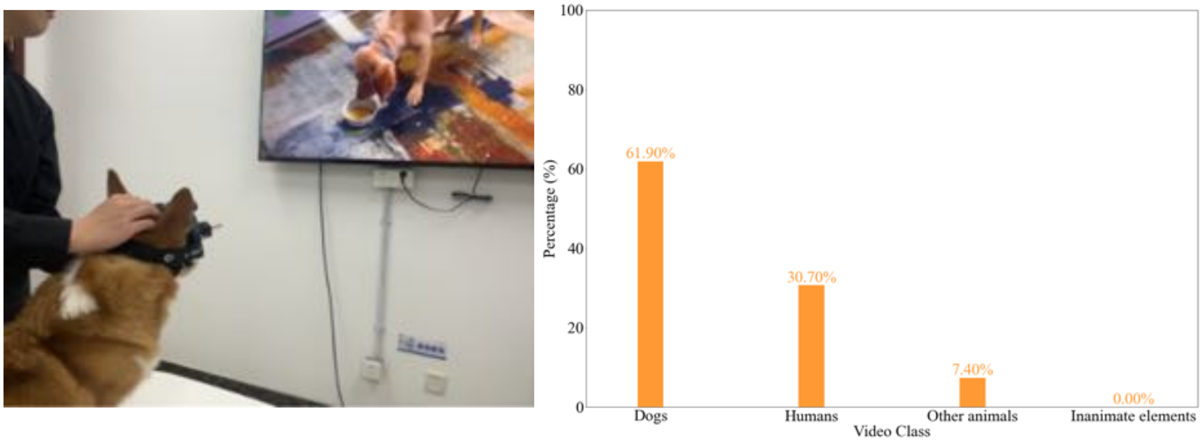


Fig. 22. *Left*: A picture of a dog is watching projected videos. *Right*: Proportion of the number of frames that attract dog visual attention to the total frame numbers in different video class.

7.1 General Discussion

In this study, eye-tracking is the primary step toward successfully uncovering causal attention in dogs. Studies have shown that dogs and humans share many aspects of the visual system. There are also works that directly apply human-based eye-tracking to dogs while using different predetermined thresholds to categorize eye movements for humans and dogs. In this work, we adopt a human-based eye-tracking method and empirically tune the hyper-parameters to obtain decent performance in recognizing dog visual attention. However, affected by many factors such as evolution and environment [74], dog eye movements differ from those of humans. There are also large differences in eye movement patterns from dog to dog, and these differences are reported to exceed those of humans. Therefore, it is valuable in the future to develop dog-specific eye-tracking approaches capable of enabling more general studies of visual systems and cognition by accommodating individual differences.

The definition of dog visual attention in this paper is based on a summary of existing literature and the assumptions of reference to human visual attention; experts might disagree with our definition. Furthermore, when designing the model, we use Gaussian relaxation to make it focus on the gaze point to intuitively represent areas of potential interest to dogs, and decay to the periphery of dogs' eye view in the form of a circle. The speed and form of this decay deserve a deep investigation. For example, literature [24, 56] states that dogs' visual acuity to the center and periphery region of gaze points is breed-dependent. These should also be further studied in the future.

The result shows that CANINE eyewear can be easily placed on dogs and used in various scenarios. As expected, the dogs can successfully perform the predesigned tasks in laboratory experiments. The five recruited dogs make counterproductive choices during the food quantity preference test in more than half of the trials. They are also adept at the signal "hand pointing", which prior research demonstrates that dogs are good at understanding and following. We also find significant individual variation when testing whether the dogs follow misleading signals from humans. For example, dog D3 consistently ignores the deceptive "throw" signal from human experimenters, while D2 and D4 obey the misleading signal most of the time (3 out of 5 trials). Although individual variation is the nature of animals, ubiquitous confounders in the real world, such as the cued objects and spatial location, may bias the study of dog behavioral responses. Our future work includes removing confounders and summarizing general measurements for studying dog cognition.

In field studies, the scenario where dogs play safeguarding roles demonstrates how dogs allocate their visual cognitive resources differently to passing humans. It is potentially valuable for studying dog territory defense, territorial aggression, etc. The quantitative index, RS, provides an additional perspective for these studies. We also use CANINE to explore how dogs and humans are attracted by different targets in similar environments, as well as investigate the scenarios where dogs interact with videos. Under these scenarios, CANINE shows its potential value. There are certainly more interesting and important scenarios to study dog cognition, e.g., how dogs interact with inanimate elements in non-social settings and how different dog characteristics, such as ages, homes, breeds, and experiences correlate to their visual cognition. It is important to generalize the model to these new/unseen scenarios. It is also important to engage the HDI research community to participate and use CANINE to expand and deepen the research landscape.

7.2 Limitations and Future Directions

A key limitation of this work is that it explores dog cognition mainly through the lens of visual stimuli. However, dogs use multiple senses to perceive the world [39]. The contributions of other senses, such as olfactory and auditory cues, to dog cognition are also important yet challenging for cognition investigation in dogs. Our future work aims to expand the CANINE algorithm and system to support the multimodal scenario to better support the exploration of dog cognition with more comprehensive capabilities.

Eye-tracking is a primary step toward exploring causal attention in dogs. Existing eye-tracking methods [42, 61] typically use an infrared camera to record the subject's eyes and use an IR LED light to illuminate the iris to detect eye-related features, such as the pupil center. The light conditions of outdoor environments can reduce the performance of such eye-tracking methods when compared to indoor environments. Our future work includes exploring methods of ameliorating the negative impacts of outdoor lighting.

8 RELATED WORK

This section reviews prior work in the three areas most closely related to our work: (1) dog cognition, (2) eye tracking on dogs, and (3) deep learning in smart glasses.

8.1 Dog Cognition

Animal cognition researchers have long studied dog cognition. In 2013, Bensky et al. comprehensively summarized the previous research on dog cognition [5]. They divided the relevant research into dog social cognition and non-social cognition. They pointed out that more than half of the studies they reviewed focused on social cognition. A recent work [3] in 2021 also pointed out that there is a growing trend in dog cognition and behavior study. In the dog social cognition research realm, Topal et al. studied dog responses to human signals in social environments [80]. Savalli suggested that dogs respond to human eye contact [10]. D'Aniello et al. indicated that dogs respond to different human visual communicative signals. Also, dogs adjust their behavior in various contexts and tasks after interpreting human communicative signals [14]. Soproni et al. studied dog responsiveness to human hand gestures [76]. Furthermore, they suggested that dogs are capable of generalizing known cues (e.g., hand pointing) to relatively new cues, such as cross-pointing and leg-pointing [76]. Lonardo et al. examined dog social cognition from another aspect. They investigated whether dogs can distinguish between true and false communicative signals and whether dogs follow misleading suggestions [49]. Similarly, Elgier et al. found that dogs responded less to deceptive cues as test trials increased. Eventually, dogs stop responding to misleading cues [9, 19, 62]. These methods typically rely on manual analysis by researchers, introducing the potential for subjective bias. This motivates the proposed automatic causal attentional inference method.

8.2 Eye Tracking on Dogs

Eye-tracking technology is able to reveal subjects' focal attention by detecting eye gaze movements [88]. Back in 2011, Williams et al. demonstrated the feasibility of applying eye-tracking to dogs [87]. They proposed a head-mounted eye-tracking camera to study visual attention in dogs during free-viewing tasks [87]. Later, Sanni et al. used an eye tracker to determine which objects attract dogs' visual attention in fixed-number object categories of pictures, which broadened the exploration of dogs' cognitive abilities [70]. They also pointed out that dogs' cognitive capacities have not yet been fully explored using eye tracking [70]. Kis et al. explored eye gaze patterns in dogs to determine when they view faces [44]. Ogura et al. investigated dog gazing behavior to reveal social visual attention using eye tracking [58]. Motivated by this trend, this work proposes to integrate eye-tracking technology into smart glasses that dogs can conveniently wear for free viewing in the open world. Coupled with the proposed deep-learning-based network, the CANINE system can automatically infer causal attention in dogs in the open world.

8.3 Deep Learning in Smart Glasses

Deep learning methods have produced transformative results and have been successfully employed in many applications [12, 38, 77, 89]. Recent work has explored the use of deep learning techniques in smart glasses to capture human cognition. For example, in 2021, Chang et al. use smart glasses MemX to detect human attention and capture human personalized interests [12]. MemX is equipped with a deep learning-based network that incorporates eye tracking and video analytics to achieve this. In emotion recognition, Zhao et al. proposed smart glasses integrating a deep-learning-based network to improve emotion recognition accuracy and understand the reasons for emotion [89]. However, deep learning techniques to access visual cognition in dogs, such as behavioral causality, are understudied. This problem requires understanding the visual scene from the dog's eye view, and further achieving reasoning. Recent advances in deep learning techniques in computer vision applications, such as Scene Graph Generation (SGG) [77] and Human Action Recognition (HAR) [38] provide a good foundation for realizing scene understanding and reasoning. SGG is a relatively new task in computer vision, which describes the objects in the scene and their relationships in the form of triples, thus supporting a large number of video-based reasoning tasks. The HAR task can identify human behavior in videos, which can potentially be used to infer dog cognition in human-dog interaction scenarios. Inspired by the latest deep learning techniques and their applications in computer vision and smart glasses, this work proposes to design a deep learning-based network to infer causal attention in dogs using smart glasses.

9 CONCLUSIONS

This work focuses on the design of algorithms and systems that can discover *causal attention* in canine vision. To overcome the obstacles faced by decades of dog cognition research, namely subjective bias and in-lab limitations, this work introduces CANINE, a deep-learning-based causal attention network and wearable system. Worn by dogs, CANINE captures dog visual attentive perceptions and extracts the rationale graph of visual attention, thereby providing data relevant to causal explanations of canine behavioral responses. The usability of the proposed wearable causal attention system has been validated through in-lab evaluations. Its generality to unconstrained real-world environments has also been illustrated through in-field trials. Using CANINE, we have collected a large-scale dataset, named DogsView, containing of automatically generated *causal attention* estimates under a wide range of representative open-world scenarios. The DogsView dataset is publicly released to support research communities that study dog cognition, human-dog interaction, dog-computer interaction, etc., and gain insights into "the world through the eyes of dogs".

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61932007 and 62090025 and in part by the Project funded by China Postdoctoral Science Foundation under Grant No. 2022T150131.

REFERENCES

[1] Bryan Agnetta, Brian Hare, and Michael Tomasello. 2000. Cues to food location that domestic dogs (Canis familiaris) of different ages do and do not use. *Animal cognition* 3, 2 (2000), 107–112.

[2] JA Araujo, ADF Chan, LL Winka, PA Seymour, and NW Milgram. 2004. Dose-specific effects of scopolamine on canine cognition: impairment of visuospatial memory, but not visuospatial discrimination. *Psychopharmacology* 175, 1 (2004), 92–98.

[3] Massimo Aria, Alessandra Alterisio, Anna Scandurra, Claudia Pinelli, and Biagio D'Aniello. 2021. The scholar's best friend: Research trends in dog cognitive and behavioral studies. *Animal Cognition* 24, 3 (2021), 541–553.

[4] Anjuli LA Barber, Dania Randi, Corsin A Müller, and Ludwig Huber. 2016. The processing of human emotional faces by pet and lab dogs: evidence for lateralization and experience effects. *PloS one* 11, 4 (2016), e0152393.

[5] Miles K Bensky, Samuel D Gosling, and David L Sinn. 2013. The world from a dog's point of view: a review and synthesis of dog cognition research. *Advances in the Study of Behavior* 45 (2013), 209–406.

[6] Michael Tomasello Brian Hare. 2005. Human-like social skills in dogs? *Trends in Cognitive Sciences* 9 (2005), 439–444.

[7] Sarah-Elizabeth Byosiere, Philippe A Chouinard, Tiffani J Howell, and Pauleen C Bennett. 2018. What do dogs (Canis familiaris) see? A review of vision in dogs and implications for cognition research. *Psychonomic bulletin & review* 25, 5 (2018), 1798–1813.

[8] Ceara Byrne, Jay Zuerndorfer, Larry Freil, Xiaochuang Han, Andrew Sirolly, Scott Cilliland, Thad Starner, and Melody Jackson. 2018. Predicting the Suitability of Service Animals Using Instrumented Dog Toys. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 127 (jan 2018), 20 pages. https://doi.org/10.1145/3161184

[9] Fabricio Carballo, Camila Cavalli, Magalí Martínez, Victoria Dzik, and Mariana Bentosela. 2020. Asking for help: Do dogs take into account prior experiences with people? *Learning & Behavior* 48, 4 (2020), 411–419.

[10] Florence Gaunet Carine Savalli, Briseida Resende. 2016. Eye Contact Is Crucial for Referential Communication in Pet Dogs. *PLOS ONE* 11, 9 (2016), e0162161.

[11] Jennifer Cattet and Ariane S Etienne. 2004. Blindfolded dogs relocate a target through path integration. *Animal Behaviour* 68, 1 (2004), 203–212.

[12] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P. Dick, Tun Lu, Ning Gu, and Li Shang. 2021. MemX: An Attention-Aware Smart Eyewear System for Personalized Moment Auto-Capture. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 2, Article 56 (June 2021), 23 pages. https://doi.org/10.1145/3463509

[13] Emma Collier-Baker, Joanne M Davis, and Thomas Suddendorf. 2004. Do dogs (Canis familiaris) understand invisible displacement? *Journal of Comparative Psychology* 118, 4 (2004), 421.

[14] Biagio D'Aniello, Anna Scandurra, Alessandra Alterisio, Paola Valsecchi, and Emanuela Prato-Previde. 2016. The importance of gestural communication: a study of human–dog communication using incongruent information. *Animal cognition* 19, 6 (2016), 1231–1235.

[15] DogsView dataset. 2022. https://github.com/MemX-Research/DogsView.

[16] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*. Computer Vision Foundation / IEEE, Long Beach, 4690–4699. https://doi.org/10.1109/CVPR.2019.00482

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] Carla Jade Eatherington, Paolo Mongillo, Miina Lõoke, and Lieta Marinelli. 2020. Dogs (Canis familiaris) recognise our faces in photographs: implications for existing and future research. *Animal cognition* 23, 4 (2020), 711–719.

[19] Angel M Elgier, Adriana Jakovcevic, Alba E Mustaca, and Mariana Bentosela. 2009. Learning and owner–stranger effects on interspecific communication in domestic dogs (Canis familiaris). *Behavioural processes* 81, 1 (2009), 44–49.

[20] C Fabrigoule and I Sagave. 1992. Reorganization of cues and path organisation in dogs. *Behavioural Processes* 28, 1-2 (1992), 65–79.

[21] Sylvain Fiset, Claude Beaulieu, Valérie LeBlanc, and Lucie Dubé. 2007. Spatial memory of domestic dogs (Canis familiaris) for hidden objects in a detour task. *Journal of Experimental Psychology: Animal Behavior Processes* 33, 4 (2007), 497.

[22] Sylvain Fiset, France Landry, and Manon Ouellette. 2006. Egocentric search for disappearing objects in domestic dogs: evidence for a geometric hypothesis of direction. *Animal Cognition* 9, 1 (2006), 1–12.

[23] Rainer Struwe Franziska Kuhne, Johanna C. Hößler. 2012. Effects of human–dog familiarity on dogs' behavioural responses to petting. *Applied Animal Behaviour Science* 142 (2012), 176–181.

[24] Márta Gácsi, Paul McGreevy, Edina Kara, and Ádám Miklósi. 2009. Effects of selection for cooperation and attention in dogs. *Behavioral and brain functions* 5, 1 (2009), 1–8.

[25] Florence Gaunet and Bertrand L Deputte. 2011. Functionally referential and intentional communication in the domestic dog: effects of spatial and social contexts. *Animal cognition* 14, 6 (2011), 849–860.

[26] Brian Hare. 2007. From nonhuman to human mind: what changed and why? *Current directions in psychological science* 16, 2 (2007), 60–64.

[27] Brian Hare, Michelle Brown, Christina Williamson, and Michael Tomasello. 2002. The domestication of social cognition in dogs. *Science* 298, 5598 (2002), 1634–1636.

[28] Brian Hare and Michael Tomasello. 1999. Domestic dogs (Canis familiaris) use human and conspecific social cues to locate hidden food. *Journal of Comparative Psychology* 113, 2 (1999), 173.

[29] Lore I Haug. 2008. Canine aggression toward unfamiliar people and dogs. *Veterinary Clinics of North America: Small Animal Practice* 38, 5 (2008), 1023–1041.

[30] Elizabeth Head, Carl W Cotman, and Norton William Milgram. 2000. Canine cognition, aging and neuropathology. *Progress in Neuro-psychopharmacology and Biological Psychiatry* 24, 5 (2000), 671–673.

[31] Ilyena Hirskyj-Douglas. 2017. *Dog Computer Interaction – Methods and Findings for Understanding how Dogs' Interact with Screens and Media*. Ph. D. Dissertation. University of Central Lancashire.

[32] Ilyena Hirskyj-Douglas and Janet C Read. 2018. DoggyVision: Examining how dogs (Canis familiaris) interact with media using a dog-driven proximity tracker device. *Animal Behavior and Cognition* 5, 4 (2018), 388–405.

[33] Ilyena Hirskyj-Douglas, Janet C Read, and Brendan Cassidy. 2017. A dog centred approach to the analysis of dogs' interactions with media on TV screens. *International Journal of Human-Computer Studies* 98 (2017), 208–220.

[34] Alexandra Horowitz. 2014. Domestic dog cognition and behavior. *The Scientific Study of Canis familiaris* (2014).

[35] Yuying Hsu and Liching Sun. 2010. Factors associated with aggressive responses in pet dogs. *Applied Animal Behaviour Science* 123, 3-4 (2010), 108–123.

[36] Björn Forkman Iben Meyer. 2015. Nonverbal Communication and Human–Dog Interaction. *Anthrozoös* 27 (2015), 553–568.

[37] Gerald H Jacobs, Jess F Deegan, Michael A Crognale, and John A Fenwick. 1993. Photopigments of dogs and foxes and their implications for canid vision. *Visual Neuroscience* 10, 1 (1993), 173–180.

[38] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10236–10247.

[39] J.L.Millot. 1994. Olfactory and visual cues in the interaction systems between dogs and children. *Behavioural Processes* 33 (1994), 177–188.

[40] Juliane Kaminski, Linda Schulz, and Michael Tomasello. 2012. How dogs know when communication is intended for them. *Developmental science* 15, 2 (2012), 222–232.

[41] Sabrina Karl, Magdalena Boch, Zsófia Virányi, Claus Lamm, and Ludwig Huber. 2020. Training pet dogs for eye-tracking and awake fMRI. *Behavior research methods* 52, 2 (2020), 838–856.

[42] Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 1151–1160.

[43] Anna Kis, Henrietta Bolló, Anna Gergely, and József Topál. 2022. Social Stimulation by the Owner Increases Dogs'(Canis familiaris) Social Susceptibility in a Food Choice Task—The Possible Effect of Endogenous Oxytocin Release. *Animals* 12, 3 (2022), 296.

[44] Anna Kis, Anna Hernádi, Bernadett Miklósi, Orsolya Kanizsár, and József Topál. 2017. The way dogs (Canis familiaris) look at human emotional faces is modulated by oxytocin. An eye-tracking study. *Frontiers in behavioral neuroscience* 11 (2017), 210.

[45] Shannon Kundey, Rebecca German, Andres De Los Reyes, Brittany Monnier, Patrick Swift, Justin Delise, and Meghan Tomlin. 2012. Domestic dogs'(Canis familiaris) choices in reference to agreement among human informants on location of food. *Animal Cognition* 15, 5 (2012), 991–997.

[46] Gabriella Lakatos, Márta Gácsi, József Topál, and Ádám Miklósi. 2012. Comprehension and utilisation of pointing gestures and gazing in dog–human communication in relatively complex situations. *Animal cognition* 15, 2 (2012), 201–213.

[47] Martina Lazzaroni, Sarah Marshall-Pescini, Helena Manzenreiter, Sarah Gosch, Lucy Přibilová, Larissa Darc, Jim McGetrick, and Friederike Range. 2020. Why do dogs look back at the human in an impossible task? Looking back behaviour may be over-interpreted. *Animal Cognition* 23, 3 (2020), 427–441.

[48] Stephen EG Lea and Britta Osthaus. 2018. In what sense are dogs special? Canine cognition in comparative context. *Learning & Behavior* 46, 4 (2018), 335–363.

[49] Lucrezia Lonardo, Christoph J Völter, Claus Lamm, and Ludwig Huber. 2021. Dogs follow human misleading suggestions more often when the informant has a false belief. *Proceedings of the Royal Society B* 288, 1955 (2021), 20210906.

[50] Sarah Marshall-Pescini, Chiara Passalacqua, Maria Elena Miletto Petrazzini, Paola Valsecchi, and Emanuela Prato-Previde. 2012. Do dogs (Canis lupus familiaris) make counterproductive choices because they are sensitive to human ostensive cues? *PLoS one* 7, 4 (2012),

e35437.

[51] Á Miklösi, Rob Polgárdi, József Topál, and Vilmos Csányi. 1998. Use of experimenter-given cues in dogs. *Animal cognition* 1, 2 (1998), 113–121.

[52] Á Miklósi, R Polgárdi, Josef Topál, and Vilmos Csányi. 2000. Intentional behaviour in dog-human communication: an experimental analysis of "showing" behaviour in the dog. *Animal cognition* 3, 3 (2000), 159–166.

[53] Ádám Miklósi, József Topál, and Vilmos Csányi. 2007. Big thoughts in small brains? Dogs as a model for understanding human social cognition. *Neuroreport* 18, 5 (2007), 467–471.

[54] Norton W Milgram, Elizabeth Head, Earl Weiner, and Earl Thomas. 1994. Cognitive functions and aging in the dog: acquisition of nonspatial visual tasks. *Behavioral neuroscience* 108, 1 (1994), 57.

[55] Norton W Milgram, Christina T Siwak, Philippe Gruet, Patricia Atkinson, Frédérique Woehrlé, and Heather Callahan. 2000. Oral administration of adrafinil improves discrimination learning in aged beagle dogs. *Pharmacology Biochemistry and Behavior* 66, 2 (2000), 301–305.

[56] Paul E Miller, Christopher J Murphy, et al. 1995. Vision in dogs. *Journal-American Veterinary Medical Association* 207 (1995), 1623–1634.

[57] Christopher J Murphy, Donald O Mutti, Karla Zadnik, and James Ver Hoeve. 1997. Effect of optical defocus on visual acuity in dogs. *American Journal of Veterinary Research* 58, 4 (1997), 414–418.

[58] Tadatoshi Ogura, Mizuki Maki, Saki Nagata, and Sanae Nakamura. 2020. Dogs (Canis Familiaris) gaze at our hands: A preliminary eye-tracker experiment on selective attention in dogs. *Animals* 10, 5 (2020), 755.

[59] Sicheng Pan, Dongsheng Li, Hansu Gu, Tun Lu, Xufang Luo, and Ning Gu. 2022. Accurate and Explainable Recommendation via Review Rationalization. In *Proceedings of the ACM Web Conference 2022*. 3092–3101.

[60] Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan kaufmann.

[61] Madeline H Pelgrim, Julia Espinosa, and Daphna Buchsbaum. 2022. Head-mounted mobile eye-tracking in the domestic dog: A new method. *Behavior Research Methods* (2022), 1–18.

[62] Mark Petter, Evanya Musolino, William A Roberts, and Mark Cole. 2009. Can dogs (Canis familiaris) detect human deception? *Behavioural Processes* 82, 2 (2009), 109–118.

[63] Oskar Pfungst. 1907. *Das Pferd des Herrn von Osten (der Kluge Hans): Ein Beitrag zur experimentellen Tier-und Menchenpsychologie.* Johann Ambrosius Barth, Leipzig.

[64] E Prato-Previde, S Marshall-Pescini, and P Valsecchi. 2008. Is your choice my choice? The owners' effect on pet dogs'(Canis lupus familiaris) performance in a food choice task. *Animal Cognition* 11, 1 (2008), 167–174.

[65] Gabriele Pretterer, Hermann Bubna-Littitz, Gerhard Windischbauer, Cornelia Gabler, and Ulrike Griebel. 2004. Brightness discrimination in the dog. *Journal of Vision* 4, 3 (2004), 10–10.

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[67] Julia Riedel, Katrin Schumann, Juliane Kaminski, Josep Call, and Michael Tomasello. 2008. The early ontogeny of human–dog communication. *Animal Behaviour* 75, 3 (2008), 1003–1014.

[68] James Serpell. 1995. *The Domestic Dog: Its Evolution, Behaviour and Interactions with People.* Cambridge University Press, Chapter 3, 21–53.

[69] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[70] Sanni Somppi, Heini Törnqvist, Laura Hänninen, Christina Krause, and Outi Vainio. 2012. Dogs do look at images: eye tracking in canine cognition research. *Animal cognition* 15, 2 (2012), 163–174.

[71] Sanni Somppi, Heini Törnqvist, Laura Hänninen, Christina M Krause, and Outi Vainio. 2014. How dogs scan familiar and inverted faces: an eye movement study. *Animal Cognition* 17, 3 (2014), 793–803.

[72] Sanni Somppi, Heini Törnqvist, Miiamaaria V Kujala, Laura Hänninen, Christina M Krause, and Outi Vainio. 2016. Dogs evaluate threatening facial expressions by their biological validity–Evidence from gazing patterns. *PloS one* 11, 1 (2016), e0143047.

[73] Sanni Somppi, Heini Törnqvist, József Topál, Aija Koskela, Laura Hänninen, Christina M Krause, and Outi Vainio. 2017. Nasal oxytocin treatment biases dogs' visual attention and emotional response toward positive human facial expressions. *Frontiers in Psychology* 8 (2017), 1854.

[74] Kenneth Holmqvist Soon Young Park, Catarina Espanca Bacelar. 2019. Dog eye movements are slower than human eye movements. *journal of eye movement research* 12 (2019).

[75] Krisztina Soproni, Ádám Miklósi, József Topál, and Vilmos Csányi. 2001. Comprehension of human communicative signs in pet dogs (Canis familiaris). *Journal of comparative psychology* 115, 2 (2001), 122.

[76] Krisztina Soproni, Ádám Miklósi, József Topál, and Vilmos Csányi. 2002. Dogs'(Canis familaris) responsiveness to human pointing gestures. *Journal of comparative psychology* 116, 1 (2002), 27.

[77] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3716–3725.

[78] Tibor Tauzin, Andor Csík, Anna Kis, and József Topál. 2015. What or where? The meaning of referential human pointing for dogs (Canis familiaris). *Journal of Comparative Psychology* 129, 4 (2015), 334.

[79] Ernő Téglás, Anna Gergely, Krisztina Kupán, Ádám Miklósi, and József Topál. 2012. Dogs' gaze following is tuned to human communicative signals. *Current Biology* 22, 3 (2012), 209–212.

[80] József Topál, Anna Kis, and Katalin Oláh. 2014. Dogs' sensitivity to human ostensive cues: a unique adaptation? In *The Social Dog*. Elsevier, 319–346.

[81] József Topál, Ádám Miklósi, Márta Gácsi, Antal Dóka, Péter Pongrácz, Enikő Kubinyi, Zsofia Viranyi, and Vilmos Csanyi. 2009. The dog as a model for understanding human social behavior. *Advances in the Study of Behavior* 39 (2009), 71–116.

[82] Heini Törnqvist, Sanni Somppi, Aija Koskela, Christina M Krause, Outi Vainio, and Miiamaaria V Kujala. 2015. Comparison of dogs and humans in visual scanning of social interaction. *Royal Society open science* 2, 9 (2015), 150341.

[83] Monique AR Udell, Robson F Giglio, and Clive DL Wynne. 2008. Domestic dogs (Canis familiaris) use human gestures but not nonhuman tokens to find hidden food. *Journal of comparative psychology* 122, 1 (2008), 84.

[84] Monique AR Udell, Kathryn Lord, Erica N Feuerbacher, and Clive DL Wynne. 2014. A dog's-eye view of canine cognition. In *domestic dog cognition and behavior*. Springer, 221–240.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Proceedings of NIPS*. 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[86] Christoph J Völter, Sabrina Karl, and Ludwig Huber. 2020. Dogs accurately track a moving object on a screen and anticipate its destination. *Scientific reports* 10, 1 (2020), 1–10.

[87] Fiona J Williams, Daniel S Mills, and Kun Guo. 2011. Development of a head-mounted, eye-tracking system for dogs. *Journal of neuroscience methods* 194, 2 (2011), 259–265.

[88] Sandra Winters, Constance Dubuc, and James P Higham. 2015. Perspectives: the looking time experimental paradigm in studies of animal visual perception and cognition. *Ethology* 121, 7 (2015), 625–640.

[89] Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. 2022. Do Smart Glasses Dream of Sentimental Visions? Deep Emotionship Analysis for Eyewear Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 1, Article 38 (mar 2022), 29 pages. https://doi.org/10.1145/3517250