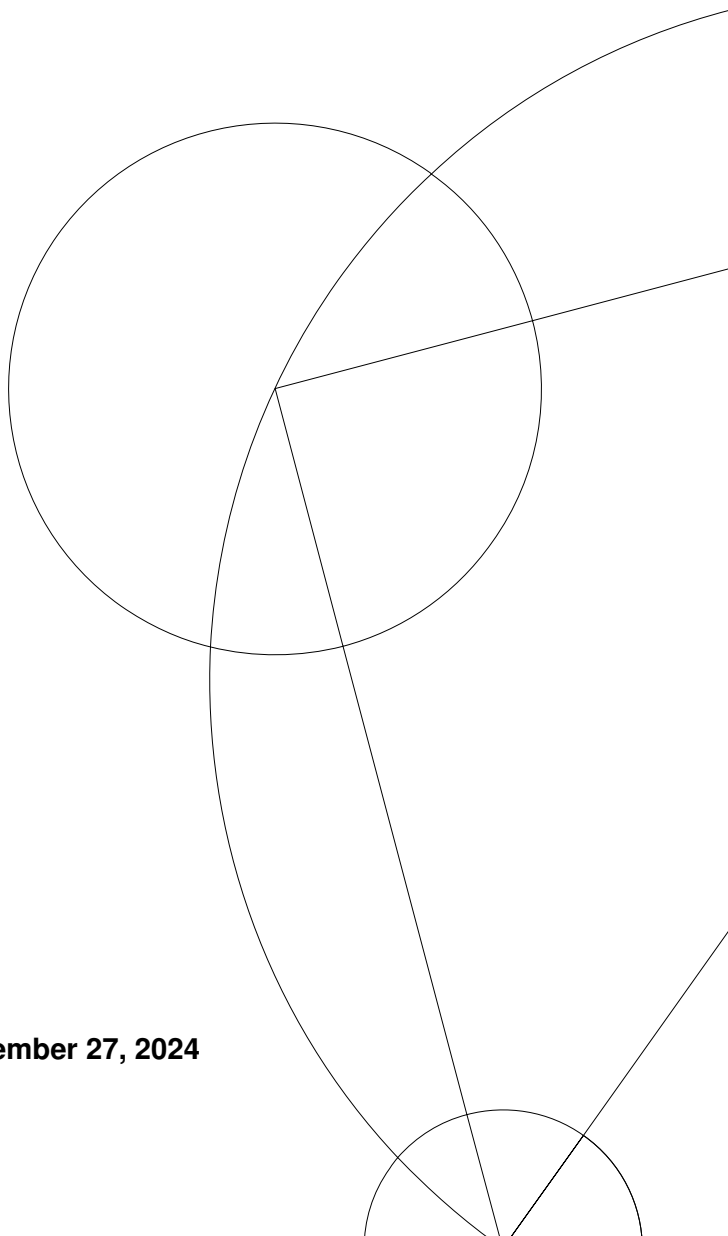
## Computational Cognitive Science III

**Group: Fri-1006040-15**

# Human vs. Machine Attention:

### Enhancing Object Detection and Segmentation with Eye-Tracking Data



**Tutor: Alberto Parola**

**Department of Nordic Studies and Languages: December 27, 2024**

## Abstract

This study investigates human and machine visual attention using the POET v1 dataset, comparing human object detection with the Faster R-CNN model and analysing attention maps generated by GradCAM and AblationCAM. Human fixation points and machine-generated attention are compared with respect to ground truth segmentation masks and further evaluated as prompts for the Segment Anything Model (SAM) in segmentation tasks. Results show that machine-generated prompts outperform human-derived points, but human attention excels in complex and occluded scenes. Despite limitations in segmentation performance, human gaze data show promise for improving automated annotation pipelines for complex images in applications such as scene understanding.

## Group Members

| Name | KUid |
|---|---|
| Magnus Lindberg Christensen | fwz302 |
| Ronja Stern | bgn595 |
| Andrea Schröter | kbn999 |

## Work Split

| Section | Main Author |
|---|---|
| Abstract | Together |
| Introduction | Together |
| Literature and Background: Human visual Attention | Ronja |
| Literature and Background: Computational Attention | Ronja |
| Literature and Background: Eye Tracking | Ronja |
| Literature and Background: Ongoing Research in Eye Tracking for Salient Object Detection | Magnus |
| Datasets | Magnus & Andrea |
| Method: Object Detection and Eye Tracking for Decision Making | Magnus |
| Method: Eye Tracking data | Andrea |
| Method: Mimicking Attentional Behavior in Machine Models | Magnus |
| Method: Comparison of Attention Points | Andrea |
| Results: Object Detection Results | Magnus |
| Results: Manual Comparison | Ronja |
| Results: Accuracy of Attention Points | Ronja |
| Results: Prompting SAM | Andrea |
| Discussion: Performance of Humans vs. Models in Object Detection | Together |
| Discussion: Comparison of human and machine generated attention points | Together |
| Conclusion | Together |

# Contents


**1 Introduction** **1**
1.1 Hypotheses and Expectations . . . 1

**2 Literature and Background** **1**
2.1 Human Visual Attention . . . 1
2.2 Computational Attention and Models . . . 2
2.3 Eye Tracking . . . 2
2.4 Ongoing Research in Eye Tracking for Salient Object Detection . . . 2

**3 Datasets** **3**
3.1 POET Dataset . . . 3
3.2 PASCAL VOC 2012 Dataset . . . 3

**4 Method** **4**
4.1 Object Detection and Eye Tracking for Decision Making . . . 4
4.2 Extracting Attention Focal Points . . . 4
4.2.1 Eye Tracking data . . . 4
4.2.2 Mimicking Attentional Behavior in Machine Models . . . 4
4.3 Comparison of Attention Points . . . 5

**5 Results** **6**
5.1 Object Detection Results . . . 6
5.2 Comparison Attention Points Results . . . 7
5.2.1 Manual Comparison . . . 7
5.2.2 Accuracy of Attention Points . . . 8
5.2.3 Prompting SAM . . . 9

**6 Discussion** **9**
6.1 Performance of Humans vs. Models in Object Detection . . . 9
6.2 Comparison of human and machine generated attention points . . . 10
6.2.1 Attention point selection . . . 10
6.2.2 Dataset constraints . . . 10
6.2.3 Proximity Matches . . . 10
6.2.4 Further Remarks . . . 10

**7 Conclusion** **10**

**A Additional Results** **13**

# 1 Introduction

Understanding visual attention is fundamental to human cognition and computer vision. Humans quickly identify relevant information in complex scenes by focusing on key areas, while machines prioritise features using computational methods. Comparing human and machine attention mechanisms is crucial for enhancing the interpretability and performance of deep learning models in computer vision. Studies have shown that aligning artificial attention with human visual attention can improve model performance in attention-driven tasks and enhance explainability in higher-level vision tasks [15]. Furthermore, understanding the differences between human and machine attention can inform the design of more robust models. For instance, while humans excel at identifying objects in complex or occluded scenes, models may struggle in such scenarios. Conversely, models might outperform humans in detecting certain categories, highlighting areas where machine attention can complement human perception. [18]

This study examines the POET v1 dataset [17], which contains images, fixation data and object detection results based on majority voting by human participants. We first compare human object detection abilities with the Faster R-CNN model, which uses a ResNet-50 FPN backbone pre-trained on COCO [28], to understand the conditions under which objects are successfully detected. Next, we analyse human fixation points using attention maps generated by GradCAM and AblationCAM. GradCAM uses gradients from the last convolutional layer of a fully convolutional network (FCN), while AblationCAM generates attention maps by systematic ablation of predictions. Local maxima extracted from these maps mimic fixation points, allowing direct comparison of human and model attention. Finally, we assess the utility of these attentional points as cues for segmentation models such as the Segment Anything Model (SAM) [14]. Eye-tracking data that identifies regions of interest (ROIs) has the potential to reduce manual annotation effort and improve segmentation model performance [36, 25].

This research aims to improve the understanding of human-machine visual attention dynamics and explore the integration of human gaze data into automated annotation pipelines. The findings underscore the importance of refining attention mechanisms to enhance both interpretability and accuracy in tasks such as object recognition and segmentation. [20, 1].

## 1.1 Hypotheses and Expectations

The following exploratory study we aim to answer the following research questions:

- Can eye-tracking data improve the process of image annotation in object detection and segmentation?
- How do human and machine attention compare in accuracy and fixation patterns?

Building on Lai et al. [15], we hypothesize: Human gaze patterns will surpass artificial models in identifying regions of interest, particularly in complex scenes with competing visual elements. By addressing this hypothesis and objectives, this study seeks to advance understanding of human-machine visual attention dynamics and consider how human gaze data can be integrated in an automated annotation pipeline.

# 2 Literature and Background

## 2.1 Human Visual Attention

When opening our eyes, huge amounts of visual information is filtered in the brain, due to lack of processing capabilities, and thus only focus on relevant details, such as regions of interest (ROIs). This is done using the mechanism of visual attention. The effectiveness of this process, inspires many computational systems to replicate it. Below, we look at various theories explaining the concept of visual attention [6].

Visual attention operates in two stages: **pre-attentive** and **attentive**. In the pre-attentive stage, basic features such as colour, intensity, orientation, motion, and edges, are rapidly extracted in parallel. These features are thus combined and integrated in the attentive stage, to identify and focus on relevant regions [37, 6]. Attention is further divided into **bottom-up** and **top-down** processes. Bottom up attention is task independent, and driven by visually salient areas, while top-down attention is guided by goals, tasks, past experiences, and mental state. This means, that the attention shifts

based on the context, and it is often found that bottom up attention is fast, while top down is slower and more deliberate [13, 24]. Finally, attention can be **overt** when eyes shifts and focuses between ROIs, searching for new information, while **covert** attentions detects changes in the peripheral vision, without applying direct focus [6, 40].

## 2.2 Computational Attention and Models

Many approaches has been made to computationally model human attention. Previous methods, like [19], extracted colour, contrast, and orientation, in a bottom up goalless manner. Though these maps may be efficient, they are less able to adapt to complex tasks compared to modern techniques.

State-of-the-art (SoTA) models, like for instance YOLO, R-CNN, mask R-CNN, RefineDet [29, 33] for object detections, and FCN, DeepLab, U-Net, and DeconvNet [21] for semantic segmentation, use convolutional neural networks (CNNs). These models learn hierarchical filters such as edges, textures, and patterns, similar to the human pre-attentive stage mentioned previously. Yet, unlike human vision, which dynamically integrates bottom-up and top-down cues from the context, CNNs lack this adaptability. This makes it difficult to understand the reasoning behind íts natural behaviour, and although SoTA models may be highly effective, they do not usually explain why certain image regions are prioritized.

Class Activation Maps (CAMs) addresses this problem by highlighting discriminative image regions from the map, and generate class specific localisation maps, [22, 42]. For instance, a Grad-CAM (Gradient-weighted Class Activation Mapping) [32] (among others) is suitable for producing semantic segmentation CAMs. It uses class specific gradient information to produce the CAM, without changing the underlying architecture [7], thus making it an applicable tool for interpretability. For object detection models, which often are not easily differentiable due to its output of for instance boxes, the gradient approach may not be suitable. Instead, gradient free approaches, such as producing e.g. an AblationCAM [5], which iteratively removes parts of the network and measures its effect, may be more suitable. Both approaches though suffer from the lack of causality. They only reveals correlations, and thus shows which areas are relevant to the output, but does not explain why or how specific features drive the predictions.

Segmentation, the process of partitioning an image into meaningful regions, benefits from advances like the Segment Anything Model (SAM) [14]. Designed to be promptable, SAM allows segmentation based on input prompts such as points or bounding boxes, enabling zero-shot transfer to new images without need for fine-tuning.

## 2.3 Eye Tracking

Eye tracking measures behavioural data, particularly the direction of an observer's gaze. Thus it provides valuable information for understanding object recognition and scene composition, as oberserver's tend to first gaze at objects of interest [34, 12]. Fixation data are commonly used in the study of visual attention [11]. Gaze direction is recorded as a point on a screen, and by measuring eye position and movement, using the *first purkinje* reflection, a light reflection from the cornea, and the pupil center, one can estimate gaze direction with reasonable accuracy. Human vision can be devided into the central 1° of vision, responsible for high-resolution detail (Fovea), the peripheral 4°, providing moderate detail (Parafovea) and the outer areas with lower resolution (Periphery). Eye-tracking data consists of points recorded per eye per frame, which can be clustered into fixations (200-300 ms periods of stable focus) and saccades (rapid movements lasing 30-80, during which vision is suppressed). Fixations can include involuntary movements (e.g. tremor, drifting, microsaccades), and saccades are often followed by brief drifts (10-40 ms), before stabilising at the next fixation [26].

## 2.4 Ongoing Research in Eye Tracking for Salient Object Detection

Papadopoulos et al. (2014) showed that fixation data can predict bounding boxes for object detectors, significantly reducing annotation time while maintaining robust performance [25]. Extending this to real-world applications, a study by Analyzing Gaze Behaviour Using Object Detection and Unsupervised Clustering applied Faster-RCNN and DBSCAN clustering to gaze data to improve the detection of visual 'looks' in natural environments, aiding fields such as social behaviour analysis [35]. These approaches highlight the potential of integrating eye-tracking data with advanced computational models for accurate and efficient object recognition.

# 3 Datasets

## 3.1 POET Dataset

We used eye-tracking data from the *POET* (Pascal Eye Tracking) v1.1 dataset [1][25], designed to identify object locations for training object detectors. The dataset is based on a subset of the *PASCAL VOC 2012* dataset (explained in subsection 3.2), and includes images across 10 object classes: *cat, dog, bicycle, motorbike, boat, aeroplane, horse, cow, sofa*, and *diningtable*. Each image is annotated with fixation points, durations, and bounding boxes retrieved from 28 participants (see Figure 1 for an example of the class *dog* with the fixation cross displayed in the eye tracking recording process and fixation points for 5 viewers.) The participants performed a *two-alternative forced choice discrimination task*, comparing paired classes (e.g., *cat/dog, bicycle/motorbike*) to avoid target-absent classes. They had to decide whether any of the objects were present in the images, as the authors argue that top-down visual search is more effective for the task than free viewing [25]. The dataset contains approximately 175.000 fixations (5.7 per viewer per image on average) with response times averaging 889 ms (ranging from 786 ms for cat to 1090 ms for cow). In total the dataset provides 6.270 PASCAL VOC trainval images along with eye-tracking annotations. Participants had unlimited time to complete tasks. [25].

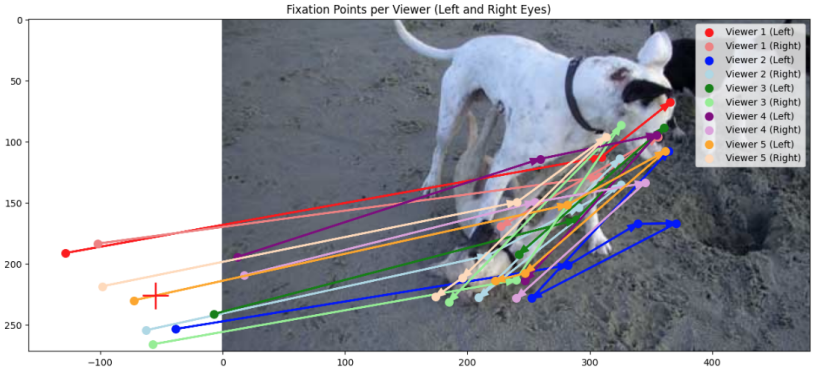



Figure 1: Example for an image in the POET data set with a fixation cross and fixation points for 5 viewers.

**Accuracy of Apparatus in the POET Dataset** The authors sat participants 60 cm from an 22' LCD screen with a 100 Hz refresh rate and 5 ms response time. Eye movements was recorded using the EyeLink 2000 eye tracker, one of the worlds most precise and accuracte eye tracker [2], while using a headrest to reduce fatigue increase precision. Button presses were recorded using a Logitech gamepad that offers millisecond accuracy. Before their studies, the authors calibrated the equipment, and performed drift correction for every 20th image, and recalibrated for every 200th image, if necessary [25].

## 3.2 PASCAL VOC 2012 Dataset

To retrieve ground truth bounding boxes and segmentation masks annotations, we accessed the *PASCAL Visual Object Classes* (VOC) 2012 dataset [3], which is commonly used as a benchmark for object detection and semantic segmentation. Unfortunately only for 1220 images of the POET dataset, segmentation were available. The dataset provides:

- 11.540 images and 27.000 object annotations
- 20 object classes: vehicles, household, animals, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, TV/monitor, bird, cat, cow, dog, horse, sheep, and person (including the subset from the POET dataset)
- Pixel-level segmentation annotations (6.929 images), bounding box annotations, and object class annotations

[1] Dataset available at `https://calvin-vision.net/datasets/poet-dataset/`
[2] `https://www.sr-research.com/eyelink-1000-plus/`
[3] Dataset available at `https://paperswithcode.com/dataset/pascal-voc`

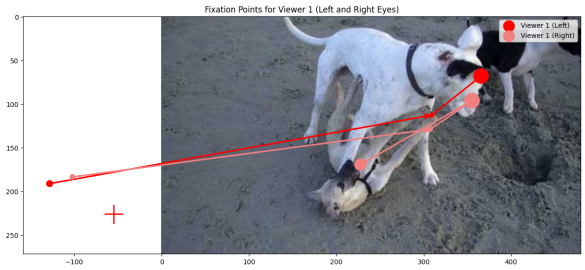



Figure 2: POET example image with fixation cross and fixation points for left and right eyes of a viewer (where greater circles indicate longer duration of fixation points).

# 4 Method

This study employs three core methods to investigate the relationship between human and machine attention in object detection and segmentation tasks. (1) We develop an automated system to detect objects in images, comparing human eye-tracking data and a pre-trained object detection model. (2) Attention points are extracted from both human participants and the model, providing a comparative basis for analyzing critical image regions relevant to object detection. (3) We systematically compare the extracted attention points using a combination of qualitative and quantitative approaches, including alignment with ground truth segmentation masks and evaluating their utility as prompts for the Segment Anything Model (SAM). The full implementation and codebase are available at https://github.com/reascr/Enhancing-Image-Segmentation-with-Eye-tracking.

## 4.1 Object Detection and Eye Tracking for Decision Making

We analyzed 6.131 images with eye-tracking data, each containing binary decisions (0 = not detected, 1 = detected) from multiple participants. Majority voting across all participants formed the decision. The agreement rate (A.R.) was calculated as the fraction of participants aligning with the majority. For machine detection, we used the Faster R-CNN model with a ResNet-50 FPN backbone pre-trained on COCO_v1 [28]. COCO has 80 classes, 123,287 images, and 886,284 instances, thus becoming larger and more complex than PASCAL VOC 2012 [27]. Since COCO includes PASCAL classes, and focuses as well on natural scenes [16], we used the model without fine-tuning, mapping COCO classes to PASCAL equivalents (e.g. sofa to couch). To align model outputs with human binary decisions (0 or 1), we needed thresholds. We tested two thresholds: (1) an IoU $> 0.5$ threshold detected once with the ground truth, and (2) an IoU $> 0.5$ threshold compared to the average IoU from comparing the max overlap between each ground truth bounding box and each predicted bounding box. This threshold is a standard for PASCAL images [30], to reduce false positives from low-probability predictions. Finally, accuracy and F1 scores were used to evaluate both the model and human performance.

## 4.2 Extracting Attention Focal Points

### 4.2.1 Eye Tracking data

To identify key points for further processing, we filtered 6,131 high-quality images from the POET dataset. We analyzed fixation patterns and excluded the first fixation point, typically centered on the fixation cross (see Figure 1). Fixation points with longer durations often corresponded to target objects (see Figure 2). Testing with SAM confirmed that using the longest-duration point as input frequently resulted in accurate segmentations of the target object. Longer fixation durations are associated with deeper cognitive processing and attention [23], as observers tend to linger on relevant objects and shift away from uninformative areas. For each observer, we selected the longest-duration fixation across both eyes, recursively choosing the next longest if the point fell outside the image. This ensured one fixation per observer, totaling five points per image. However, this selection remains heuristic and not a a reliable method for selecting the most informative fixation point.

### 4.2.2 Mimicking Attentional Behavior in Machine Models

In our study, participants performed a top-down task. CAMs were applied to two models in different variations: (1) a fully convolutional network (FCN) with a ResNet-50 backbone, pretrained on PAS-

CAL data [4]. We used a GradCAM [31] to mimick the model attention by predicting using the last convolutional layer *layer4* in the model *backbone*. For (2) object detection we had to use a gradient free approach, and therefore used the AblationCAM discussed earlier [5]. Here, we set the detection threshold to 0.0 to produce a complete attention map, and used the model *backbone* as the target layer for prediction. Both models drew on approaches from [8, 7]. Examples of each cam can be seen in Figure 3.

(a) Original Image (b) AblationCAM (c) GradCAM

Figure 3: An example of the object detection AblationCAM and the semantic segmentation GradCAM for the class *horse* used to derive the points for prompting

To identify attention points, we extract local maxima from the heat map, as higher values represent regions of greater attention. These maxima must be at least 10 pixels apart to ensure spatial separation. A threshold of 0.7 is applied to exclude low intensity peaks and prevent irrelevant areas from being selected as attention points. The extracted maxima are then sorted by intensity, prioritising the most prominent attentional regions. If no points are detected under these conditions, the threshold is relaxed to 0 to ensure that attentional points can still be identified.

### 4.3 Comparison of Attention Points

We investigate where attention points of different models tend to cluster by **manually reviewing images** and their corresponding differences. This qualitative analysis aims to observe any obvious patterns, such as whether certain regions of an image consistently attract attention in one model but not the other. While this evaluation does not yield quantitative metrics, it provides an intuitive understanding of how attention points are distributed across models. Studies highlight that aligning artificial attention with human gaze patterns can improve model performance in attention-driven tasks and enhance explainability in higher-level vision applications. Human gaze patterns naturally focus on salient features, making them a valuable benchmark for evaluating machine attention mechanisms. [15]

Next, we assess the **accuracy of attention points** by determining how many fall within the ground truth segmentation mask. To refine this evaluation, we include proximity matches, considering points near the mask boundary as valid. This approach acknowledges attention to adjacent regions, such as borders or edges, which may aid object recognition. Proximity matches also account for scenarios where a human observer focuses on a point outside the mask but uses peripheral vision to identify the object. Peripheral vision significantly contributes to implicit attentional learning, supporting a nonselective pathway that guides visual search [2].

Finally, we evaluate whether attention points serve as **effective prompts for the Segment Anything Model** (SAM). This analysis examines which points—fixation-based or model-generated—yield better segmentations when used as prompts. Studies indicate that segmentation quality can vary significantly based on the type of prompt used; therefore, we aim to investigate differences in segmentation outcomes [3]. The segmentation quality, primarily influenced by SAM's capabilities, is assessed through generated segmentation masks. These masks (see Figure 4), along with the ground truth, are binarized, and the Dice coefficient—a measure of overlap between binary masks—is calculated for each image. To assess whether there is a significant performance difference between the human and machine-generated points, a Wilcoxon signed-rank test was performed [38].

$$\text{Dice coefficient} = \frac{2|A \cap B|}{|A| + |B|}, \text{where A and B are binarized masks.}$$

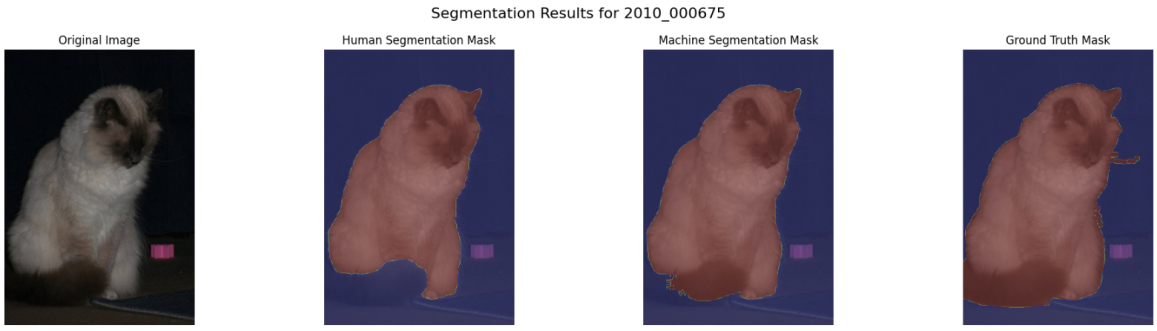



Figure 4: Example of human and machine segmentation masks after passing 5 points to SAM (where the machine segmentation mask was generated based on Ablation model points).

For both tasks "Accuracy of Attention Points" and "Prompts for SAM" we are using using 1220 images where we have both fixation data from the POET v1 dataset [17] and ground truth segmentation masks from the Pascal VOC 2012 dataset. The images were further categorised into four groups:

- Images where both the model and human observers successfully detected the object.
- Images where only human observers were successful.
- Images where only the model was successful.
- Images in which both were unsuccessful.

To ensure fairness, comparisons were made using either a single fixation point compared to a single heatmap maximum, or a set of five fixation points compared to five extracted heatmap maximums. For computational models the point with highest confidence was extracted for one point comparisons. There were cases where fixation points were missing (e.g. outside the image) or fewer than five maxima could be extracted from the heatmap (e.g. due to computational limitations or insufficient distinct maxima). In cases of missing points, a False value was assigned when evaluating whether the points fell within the segmentation mask. Similarly, if fewer than five points were available, the remaining points were still used as input to SAM. For images with no available points, a Boolean mask containing only False values was used to compute the Dice coefficient relative to the ground truth mask. Additionally, images were filtered to ensure that model-generated attention points were available for comparison (in case of SAM).

# 5 Results

## 5.1 Object Detection Results

The results for the object detection task show that the best performing model is the IoU $> 0.5$ single, but this also has better prerequisites for detection than the IoU $> 0.5$ average. Looking at the humans overall, they generally perform better than any of the models, yet their overall accuracy and f1 score falls short of those from the model. The inter viewer variability shows the there a large differences in the individual performances. Viewer 1 and Viewer 2 show almost perfect performance, while viewer 3, 4, and 5 struggles in some of the categories, like for *dog*, *cat*, *cow* and *sofa*.

| Case | Metric | Cat | Dog | Bicycle | Motorbike | Boat | Aero | Horse | Cow | Sofa | D.Table | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Human Eye Tracking** | | | | | | | | | | | | |
| **Overall** | **Acc.** | 0.695 | 0.691 | **0.996** | **0.990** | **0.995** | **0.997** | **1.000** | **0.989** | 0.976 | **0.991** | 0.885 |
| **Overall** | **F1** | 0.820 | 0.817 | 0.998 | 0.995 | 0.998 | 0.998 | 1.000 | 0.995 | 0.988 | 0.996 | 0.939 |
| **Viewer 1** | **Acc.** | 0.978 | 0.982 | 0.950 | 0.943 | 0.965 | 0.994 | 0.983 | 0.875 | 0.942 | 0.965 | 0.966 |
| **Viewer 2** | **Acc.** | 0.982 | 0.984 | 0.980 | 0.933 | 0.936 | 0.969 | 0.991 | 0.983 | 0.972 | 0.985 | 0.974 |
| **Viewer 3** | **Acc.** | 0.666 | 0.644 | 0.984 | 0.974 | 0.890 | 0.934 | 0.987 | 0.956 | 0.872 | 0.945 | 0.839 |
| **Viewer 4** | **Acc.** | 0.613 | 0.634 | 0.979 | 0.967 | 0.986 | 0.989 | 0.971 | 0.849 | 0.968 | 0.982 | 0.845 |
| **Viewer 5** | **Acc.** | 0.685 | 0.680 | 0.981 | 0.975 | 0.965 | 0.974 | 0.979 | 0.966 | 0.923 | 0.957 | 0.864 |
| **FasterRCNN Model Variations, Gtbb True** | | | | | | | | | | | | |
| **M1: IoU > 0.5, Single** | **Acc.** | **0.986** | **0.987** | 0.980 | 0.984 | 0.959 | 0.992 | 0.992 | 0.983 | **0.981** | 0.984 | **0.983** |
| **M1: IoU > 0.5, Single** | **F1** | 0.993 | 0.993 | 0.990 | 0.992 | 0.979 | 0.996 | 0.995 | 0.992 | 0.990 | 0.992 | 0.992 |
| **M2: IoU > 0.5, Average** | **Acc.** | 0.981 | 0.977 | 0.962 | 0.976 | 0.924 | 0.968 | 0.983 | 0.963 | 0.979 | 0.951 | 0.969 |
| **M2: IoU > 0.5, Average** | **F1** | 0.991 | 0.989 | 0.980 | 0.988 | 0.961 | 0.984 | 0.992 | 0.981 | 0.989 | 0.975 | 0.984 |
| **Overall Human vs. Models** | | | | | | | | | | | | |
| **M1** | **A.R.** | 0.691 | 0.686 | 0.977 | 0.978 | 0.955 | **0.992** | 0.991 | 0.980 | 0.961 | 0.976 | 0.873 |
| **M2** | **A.R.** | 0.690 | 0.684 | 0.958 | 0.971 | 0.920 | 0.971 | **0.983** | 0.960 | 0.959 | 0.943 | 0.861 |

Table 1: FasterRCNN model variations performance with a ground truth bounding box (Gtbb) IoU threshold, compared to human detection performance. The table shows the detection accuracy (Acc.), F1 score (F1), and Agreement Ratio (A.R.). Computed across all available images containing sufficient eye tracking data. Highest values emphasized in **Bold**, lowest values emphasized with underline.

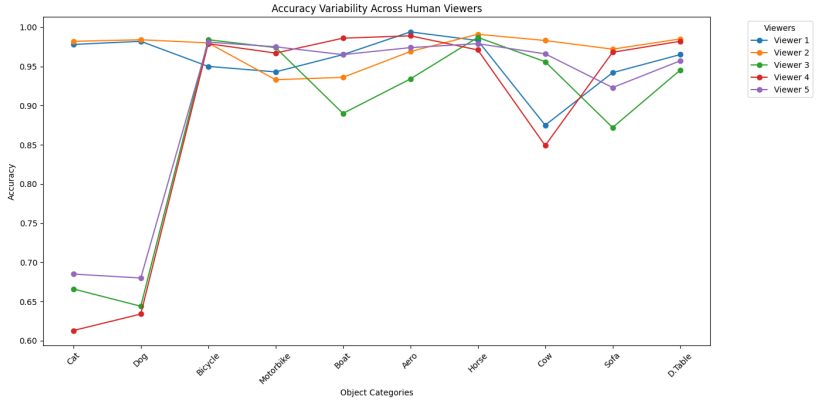


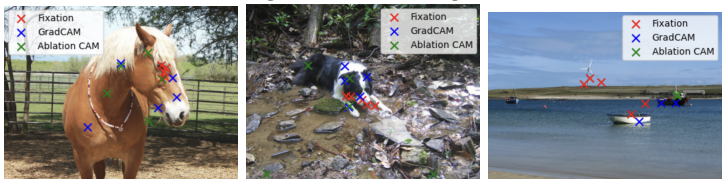

Figure 5: Accuracy Variability Across All Participants

## 5.2 Comparison Attention Points Results

### 5.2.1 Manual Comparison

Example images with extracted attention points are shown for the four subsets. The "both right" subset, with 1,033 images the biggest subset, "Human Right" with 32 images, "Model Right" with 150 images, and "Both Wrong" with only 5 images. In particular, the "Model Right" subset has a high proportion of images from the Cat and Dog classes.

Figure 6: Both Right

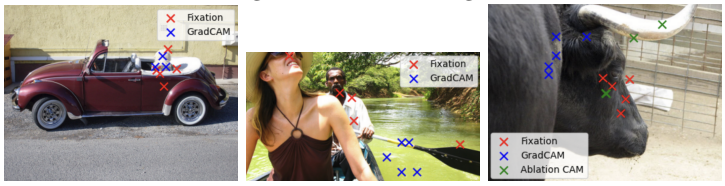



(a) Horse (b) Dog (c) Boat

Figure 7: Both Wrong

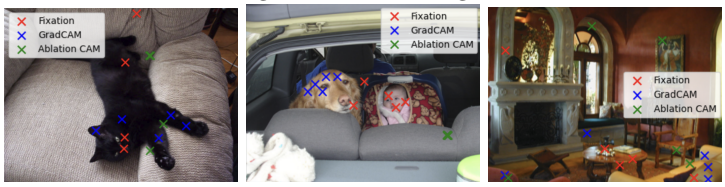

(a) Cat (b) Cat (c) Dog

Figure 8: Human Right




(a) Car (b) Boat (c) Cow

Figure 9: Model Right




(a) Cat (b) Dog (c) Sofa

### 5.2.2 Accuracy of Attention Points

| Split | Length | GradCAM | Ablation | Viewer 1 | Viewer 2 | Viewer 3 | Viewer 4 | Viewer 5 |
|---|---|---|---|---|---|---|---|---|
| **Exact Match** | | | | | | | | |
| Both Right | 1033 | **0.860** | 0.637 | 0.676 | 0.635 | 0.580 | 0.607 | 0.661 |
| Human Right | 32 | **0.531** | 0.250 | 0.500 | 0.375 | 0.500 | 0.406 | 0.500 |
| Model Right | 150 | **0.973** | 0.767 | 0.813 | 0.720 | 0.847 | 0.753 | 0.773 |
| Both Wrong | 5 | 0.600 | 0 | **0.800** | 0.400 | 0.400 | 0.400 | 0.600 |
| **Proximity Match (10 pixels)** | | | | | | | | |
| Both Right | 1033 | **0.951** | 0.760 | 0.829 | 0.796 | 0.744 | 0.774 | 0.847 |
| Human Right | 32 | 0.656 | 0.344 | **0.688** | 0.594 | 0.656 | 0.562 | 0.750 |
| Model Right | 150 | **0.973** | 0.820 | 0.900 | 0.807 | 0.887 | 0.887 | 0.873 |
| Both Wrong | 5 | 0.800 | 0 | **1** | 0.600 | 0.600 | **1** | **1** |

Figure 11: Accuracy scores across 1 attention points indicating whether they lie in or outside of the segmentation mask, best scores in **bold**


Figure 12: Attention Points for the class Bike

GradCAM attention points consistently show the highest accuracy for both 1 and 5 attention points, with human fixation points only outperforming them in proximity matches for images where the model failed to detect the object (human right and both wrong). Ablation attention points and fixation points have similar scores, except in cases where the model failed, resulting in near-zero scores.

In Figure 12 the attention points for a bike image are visualised. Detailed accuracy values for each object class are shown in the appendix (Table 2). For the subset where both human and model predictions are correct, the bicycle class

| Split | Length | GradCAM | Ablation | Fixation |
|---|---|---|---|---|
| **Exact Match** | | | | |
| Both Right | 1033 | **0.867** | 0.640 | 0.632 |
| Human Right | 32 | **0.550** | 0.325 | 0.456 |
| Model Right | 150 | **0.973** | 0.751 | 0.781 |
| Both Wrong | 5 | **0.680** | 0.040 | 0.520 |
| **Proximity Match (10 pixels)** | | | | |
| Both Right | 1048 | **0.943** | 0.761 | 0.798 |
| Human Right | 32 | 0.619 | 0.400 | **0.650** |
| Model Right | 150 | **0.979** | 0.805 | 0.871 |
| Both Wrong | 5 | 0.720 | 0.040 | **0.840** |

Figure 10: Accuracy scores across 5 attention points indicating whether they lie in or outside of the segmentation mask, best scores in **bold**

has the lowest exact match scores: 0.277 for GradCAM and 0.255 for fixation. However, using close matches, these scores improve significantly to 0.824 for GradCAM and 0.699 for fixation, marking the largest performance gains of all classes.

#### 5.2.3 Prompting SAM

Figure 13 shows the mean Dice coefficients for comparing segmentation masks generated using human and machine-generated attention points (GradCAM and Ablation) with ground truth segmentation masks from the PASCAL VOC dataset. The "Both Wrong" condition was excluded due to insufficient data (only 5 images for GradCAM and 1 for Ablation points).

Generally, machine-generated attention points outperform human fixation points. GradCAM points consistently achieved higher Dice coefficients than Ablation points, especially in the "Both Right" and "Model Right" conditions. However, in the "Human Right" condition, where the model failed to detect objects, human fixation points yielded better segmentations than GradCAM, and Ablation model generated points outperformed GradCAM generated points. Passing SAM five points also led to higher Dice coefficients overall.

| | | 5 Points Prompt | | 1 Point Prompt | |
|---|---|---|---|---|---|
| **Filter** | **Length** | **Model** | **DICE** | **Model** | **DICE** |
| Both right | 1033 | Human | 0.580 | Viewer 1 | 0.335 |
| | | | | Viewer 2 | 0.336 |
| | | | | Viewer 3 | 0.306 |
| | | | | Viewer 4 | 0.321 |
| | | | | Viewer 5 | 0.327 |
| | 1033 | GradCAM | **0.661** | GradCAM | **0.426** |
| | 1033 | Ablation | 0.610 | Ablation | 0.355 |
| Human Right | 32 | Human | **0.555** | Viewer 1 | 0.299 |
| | | | | Viewer 2 | 0.217 |
| | | | | Viewer 3 | **0.352** |
| | | | | Viewer 4 | 0.307 |
| | | | | Viewer 5 | 0.291 |
| | 32 | GradCAM | 0.329 | GradCAM | 0.173 |
| | 23 | Ablation | 0.437 | Ablation | 0.207 |
| Model Right | 150 | Human | 0.628 | Viewer 1 | 0.524 |
| | | | | Viewer 2 | 0.477 |
| | | | | Viewer 3 | 0.547 |
| | | | | Viewer 4 | 0.464 |
| | | | | Viewer 5 | 0.522 |
| | 150 | GradCAM | **0.698** | GradCAM | **0.571** |
| | 150 | Ablation | 0.689 | Ablation | 0.519 |

Figure 13: Mean Dice Coefficients for comparing segmentation masks generated by human and machine attention points (GradCAM and Ablation) with ground truth segmentation masks from PASCAL VOC. Highest values are emphasized in bold.

Statistical analysis using Wilcoxon signed-rank tests supports these results. In the "Both Right" condition, both GradCAM ($p = 4.3849e\text{-}13$) and Ablation ($p = 3.6430e\text{-}05$) significantly outperformed human fixation points. In the "Human Right" condition, human fixation points outperformed GradCAM points ($p = 0.0017$), while no significant difference was found between Ablation points and human fixation points ($p = 0.3931$). In the "Model Right" condition, both GradCAM ($p = 0.0001$) and Ablation ($p = 0.0034$) significantly outperformed human fixation points.

# 6 Discussion

## 6.1 Performance of Humans vs. Models in Object Detection

While humans generally outperformed the FasterRCNN models in object detection, they showed lower accuracy for the Cat and Dog classes. Despite accurately locating fixation points within the segmentation masks (see Appendix Table 2), humans struggled with classification, particularly for these two classes, which may reflect incomplete attention or misclassification (as evidenced by shorter fixation times for cats). Human performance was also more variable due to individual differences, distraction and noise in the eye-tracking data, consistent with previous findings [9, 41].

Further, the choice of threshold was key to valid detections. We used a detection threshold of 0.0 and an IoU threshold of 0.5, balancing precision and false positives, chosen based on heuristics from prior research [30]. Alternative thresholds were not explored, therefore, future work should evaluate broader IoU thresholds to improve detection accuracy and localisation.

Lastly, we did not fine-tune the FastRCNN model on PASCAL, as it was pre-trained on COCO to balance efficiency and performance. While fine-tuning could reduce misclassification errors, it had minimal impact on the model's performance in comparing attention mechanisms.

### 6.2 Comparison of human and machine generated attention points

GradCAM generally outperformed both AblationCAM and human fixation points in most conditions, especially "Both Right" and "Model Right". GradCAM outperformed AblationCAM, probably due to its semantic segmentation base, which focuses on pixel-level attention. In contrast, AblationCAM, derived from an object detection model, emphasised larger object areas. However, in Human Right and Both Wrong, human fixation points provided better segmentations and were more consistent within the mask. This suggests that human attention may be better suited to images that require contextual reasoning, such as those where subtle cues (e.g. a boat inferred from a paddle and water Figure 8 **c**) or object occlusion (e.g. a cat partially visible behind a fence Figure 7 **b**) are important.

#### 6.2.1 Attention point selection

Choosing fixation points based on longest duration was crucial, due to linking longer fixations to deeper cognitive processing [23], it was not always optimal. Although this method is effective in identifying deeper cognitive processing, it may not always coincide with the target object. Future research should refine the selection of fixation points to improve segmentation.

#### 6.2.2 Dataset constraints

The dataset had significant variability in fixation points, raising concerns about generalisation. Limited ground truth segmentation masks and the fact that only a small and inconsistent subset of participants viewed each image (5 out of 28) introduced potential bias, particularly for classes such as dog and cat. Participant fatigue also likely affected data reliability and segmentation accuracy. Lastly, the task design (focusing on object presence rather than fixation), which shapes top-down attention, likely influenced the results by reducing the precision of fixation points and their effectiveness in guiding segmentation.

#### 6.2.3 Proximity Matches

Proximity matching significantly improved accuracy for all attentional points, with the greatest gain for fixation points. This highlights the importance of boundary placement, as fixation points are often close to object boundaries. This suggests, that that viewers use peripheral vision and scene understanding to detect objects [39]. Individual differences, distraction and noise in eye-tracking data - such as calibration errors and head movements [10] - can reduce performance, especially when SAM is used.

#### 6.2.4 Further Remarks

Bikes posed a challenge for segmentation, with fixation points often outside the segmentation mask (see Figure 12). Proximity matching significantly improved accuracy more than for any other class from 0.25 and 0.28 to 0.69 and 0.83, respectively for both fixation and GradCAM points for bikes. In addition, using five points instead of one to prompt SAM improved segmentation, suggesting that multiple attention points provide complementary information. Human fixation points, while capable of capturing finer detail, require more data and a well-designed task to overcome individual variability and improve segmentation performance.

## 7 Conclusion

This study investigated human and model performance in an object detection task, compared human fixation points with machine-generated attention maps from GradCAM and AblationCAM, and explored the integration of human eye-tracking data for segmentation tasks using SAM. Contrary to our expectations, machine-generated cues generally led to higher accuracy. While humans struggled to detect object classes such as dogs and cats, their fixation points still remain in salient regions, suggesting potential for segmentation tasks. Human fixations outperformed in certain cases, particularly when contextual reasoning was required, highlighting the importance of understanding when and why each approach outperforms. The variability in human performance highlights the need for better task design and participant assessment. Overall, while human fixation data alone does not outperform machine performance, it shows stronger performance in complex scenes and could complement models in certain scenarios with further refinement.

# References


[1] P. Banerjee, S. Raj, and D. R. P. Barnwal. Machine versus human attention: A comparative study of different transfer learning models. In *Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD)*, CODS-COMAD '23, page 136, New York, NY, USA, 2023. Association for Computing Machinery.

[2] C. Chen and V. G. Lee. Peripheral vision contributes to implicit attentional learning: Findings from the "mouse-eye" paradigm. *Attention, Perception Psychophysics*, 86(5):1621–1640, Jul 2024.

[3] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li. Sam on medical images: A comprehensive study on three prompt modes, 2023.

[4] T. Contributors. *FCN ResNet50 Model in TorchVision*, 2024. Accessed: 2024-12-11.

[5] s. desai and H. G. Ramaswamy. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, March 2020.

[6] M. S. Gide and L. J. Karam. *Computational Visual Attention Models*. Now Publishers, 2017. Available via ProQuest Ebook Central.

[7] J. Gil. Class activation maps for object detection with faster rcnn, 2024. Accessed: 2024-12-11.

[8] J. Gil. Class activation maps for semantic segmentation, 2024. Accessed: 2024-12-11.

[9] T. Harada, H. Iwasaki, K. Mori, A. Yoshizawa, and F. Mizoguchi. Evaluation model of cognitive distraction state based on eye-tracking data using neural networks. In *2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing*, pages 428–434, 2013.

[10] K. Holmqvist and R. Andersson. Eye tracking: A comprehensive guide to methods. *Paradigms and measures*, 2017.

[11] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10):1489–1506, 2000.

[12] L. Itti and C. Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[13] F. Katsuki and C. Constantinidis. Bottom-up and top-down attention: different processes and overlapping neural systems. *The Neuroscientist*, 20(5):509–521, October 2014. Published online 2013 Dec 20.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[15] Q. Lai, S. Khan, Y. Nie, J. Shen, H. Sun, and L. Shao. Understanding more about human and machine attention in deep neural networks, 2020.

[16] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft coco: Common objects in context, 2015.

[17] A. X. C. A. H. H. D. I. A. F. S. S. Manolis Savva, Luca Weihs. Poet: Physical object exploration and tracking dataset. `https://calvin-vision.net/datasets/poet-dataset/`, 2024. Accessed: 2024-12-16.

[18] M. McDonough. The limits of computer vision, and of our own. *Harvard Medical School Magazine*, 2024. Feature article, 13 min read, Spring 2024.

[19] O. L. Meur, P. L. Callet, D. Barba, and D. Thoreau. A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:802–817, 2006.

[20] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

[21] Y. Mo, Y. Wu, X. Yang, F. Liu, and Y. Liao. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646, 2022.

[22] M. A. Morid, A. Borjali, and G. D. Fiol. A scoping review of transfer learning research on medical image analysis using ImageNet. *Computers in Biology and Medicine*, 128:104115, 2021.

[23] S. Negi and R. Mitra. Fixation duration and the learning process: an eye tracking study with subtitled videos. *Journal of Eye Movement Research*, 13(6):10.16910/jemr.13.6.1, 2020.

[24] Z. Niu, G. Zhong, and H. Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[25] D. P. Papadopoulos, A. D. Clarke, F. Keller, and V. Ferrari. Training object class detectors from eye tracking data. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 361–376. Springer, 2014.

[26] A. Parola. Introduction to computational cognitive science 3. Lecture Presentation, September 2024. Presented at the University of Copenhagen.

[27] J. Pont-Tuset and L. Van Gool. Boosting object proposals: From pascal to coco. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[28] PyTorch. *torchvision.models.detection.fasterrcnn_resnet50_fpn*, 2024. Accessed: 2024-12-03.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[30] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese. Generalized intersection over union: A metric and a loss for bounding box regression, 2019.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2019.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, Feb. 2020.

[33] F. Sultana, A. Sufian, and P. Dutta. *A Review of Object Detection Models Based on Convolutional Neural Network*, pages 1–16. Springer Singapore, Singapore, 2020.

[34] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980.

[35] P. Venuprasad, L. Xu, E. Huang, A. Gilman, L. C. Ph.D., and P. Cosman. Analyzing gaze behavior using object detection and unsupervised clustering. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Full Papers, New York, NY, USA, 2020. Association for Computing Machinery.

[36] B. Wang, A. Aboah, Z. Zhang, and U. Bagci. Gazesam: What you see is what you segment. *arXiv preprint arXiv:2304.13844*, 2023.

[37] J. M. Wolfe and I. S. Utochkin. What is a preattentive feature? *Current Opinion in Psychology*, 29:19–26, October 2019. Epub 2018 Nov 13.

[38] R. F. Woolson. Wilcoxon signed-rank test. *Encyclopedia of Biostatistics*, 8, 2005.

[39] E. Wästlund, P. Shams, and T. Otterbring. Unsold is unseen ... or is it? examining the role of peripheral vision in the consumer choice process using eye-tracking methodology. *Appetite*, 120:49–56, 2018.

[40] D. Zanca, M. Gori, S. Melacci, and A. Rufa. Gravitational models explain shifts on human visual attention. *Scientific Reports*, 10(1):16335, October 2020. Epub 2020 Oct 1.

[41] H. Zhang, M. R. H. Smith, and G. J. Witt. Identification of real-time diagnostic measures of visual distraction with an automatic eye-tracking system. *Human Factors*, 48(4):805–821, 2006. PMID: 17240726.

[42] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.

# Appendix

# A Additional Results

| **Match** | **Class** | **Both Right** | | | **Human Right** | | | **Model Right** | | | **Both Wrong** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | GradCAM | Ablation | Fix_Mean | GradCAM | Ablation | Fix_Mean | GradCAM | Ablation | Fix_Mean | GradCAM | Ablation | Fix_Mean |
| Exact | aeroplane | 0.913 | 0.722 | 0.574 | 1 | 0.4 | 0.8 | - | - | - | - | - | - |
| Exact | horse | 0.964 | 0.79 | 0.642 | 0.5 | 0.7 | 0.3 | - | - | - | - | - | - |
| Exact | cat | 0.995 | 0.785 | 0.875 | 0 | 0.5 | 0.5 | 0.997 | 0.778 | 0.842 | 0.667 | 0.067 | 0.667 |
| Exact | sofa | 0.957 | 0.676 | 0.559 | 0.333 | 0 | 0.667 | 0.7 | 0.3 | 0.1 | - | - | - |
| Exact | cow | 0.937 | 0.846 | 0.716 | 1 | 0.533 | 0.667 | 0.9 | 0.6 | 0.3 | - | - | - |
| Exact | motorbike | 0.87 | 0.612 | 0.709 | - | - | - | 0.2 | 0 | 0.2 | 0.4 | 0 | 0 |
| Exact | dog | 0.983 | 0.756 | 0.803 | 0.9 | 0.45 | 0.65 | 0.977 | 0.771 | 0.786 | 1 | 0 | 0.6 |
| Exact | diningtable | 0.896 | 0.258 | 0.529 | 0.9 | 0 | 0.5 | 1 | 0 | 0.2 | - | - | - |
| Exact | boat | 0.84 | 0.604 | 0.55 | 0.571 | 0.457 | 0.343 | 1 | 1 | 1 | - | - | - |
| Exact | bicycle | **0.277** | **0.188** | **0.255** | 0.033 | 0.033 | 0.1 | - | - | - | - | - | - |
| Proximity | aeroplane | 0.961 | 0.893 | 0.777 | 1 | 0.533 | 0.933 | - | - | - | - | - | - |
| Proximity | horse | 0.97 | 0.901 | 0.828 | 0.5 | 0.7 | 0.6 | - | - | - | - | - | - |
| Proximity | cat | 0.997 | 0.832 | 0.924 | 0 | 0.5 | 0.8 | 1 | 0.842 | 0.919 | 0.667 | 0.067 | 0.867 |
| Proximity | sofa | 0.957 | 0.711 | 0.651 | 0.333 | 0 | 0.733 | 0.75 | 0.4 | 0.15 | - | - | - |
| Proximity | cow | 0.943 | 0.931 | 0.863 | 1 | 0.6 | 0.867 | 0.9 | 1 | 0.5 | - | - | - |
| Proximity | motorbike | 0.939 | 0.737 | 0.854 | - | - | - | 0.6 | 0.2 | 1 | 0.6 | 0 | 1 |
| Proximity | dog | 0.987 | 0.807 | 0.897 | 0.9 | 0.6 | 0.85 | 0.977 | 0.803 | 0.878 | 1 | 0 | 0.6 |
| Proximity | diningtable | 0.915 | 0.301 | 0.644 | 0.9 | 0 | 0.6 | 1 | 0 | 0.2 | - | - | - |
| Proximity | boat | 0.902 | 0.733 | 0.735 | 0.571 | 0.514 | 0.457 | 1 | 1 | 1 | - | - | -n |
| Proximity | bicycle | **0.824** | **0.568** | **0.699** | 0.4 | 0.167 | 0.433 | - | - | - | - | - | - |

Table 2: Evaluation results for GradCAM, Ablation, and Fix_Mean for the different object classes