Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention

Sounak Mondal[1], Zhibo Yang[1,2], Seoyoung Ahn[1], Dimitris Samaras[1], Gregory Zelinsky[1], Minh Hoai[1,3]
[1]Stony Brook University, [2]Waymo LLC, [3]VinAI Research

Abstract

Predicting human gaze is important in Human-Computer Interaction (HCI). However, to practically serve HCI applications, gaze prediction models must be scalable, fast, and accurate in their spatial and temporal gaze predictions. Recent scanpath prediction models focus on goal-directed attention (search). Such models are limited in their application due to a common approach relying on trained target detectors for all possible objects, and the availability of human gaze data for their training (both not scalable). In response, we pose a new task called ZeroGaze, a new variant of zero-shot learning where gaze is predicted for never-before-searched objects, and we develop a novel model, Gazeformer, to solve the ZeroGaze problem. In contrast to existing methods using object detector modules, Gazeformer encodes the target using a natural language model, thus leveraging semantic similarities in scanpath prediction. We use a transformer-based encoder-decoder architecture because transformers are particularly useful for generating contextual representations. Gazeformer surpasses other models by a large margin (19%–70%) on the ZeroGaze setting. It also outperforms existing target-detection models on standard gaze prediction for both target-present and target-absent search tasks. In addition to its improved performance, Gazeformer is more than five times faster than the state-of-the-art target-present visual search model. Code can be found at https://github.com/cvlab-stonybrook/Gazeformer/

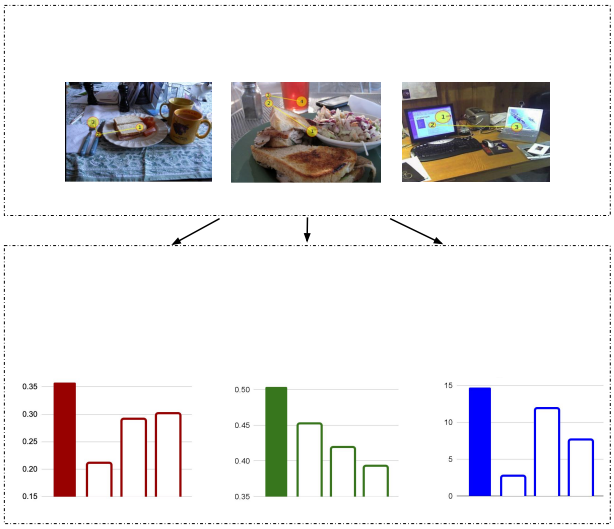

Figure 1. An example scenario where human scanpaths for only N categories such as "fork", "cup" and "laptop" are available during training. In the traditional *GazeTrain* setting, the model is asked to predict scanpaths for only those N target categories it has seen during training, such as "fork". However, in the *ZeroGaze* setting, the model is expected to predict scanpaths for target categories for which training scanpaths are *not available*, such as "knife". Gazeformer is overall superior since it is more scalable (better ZeroGaze performance), more effective (better GazeTrain performance) and faster (higher inference speed) than previous methods.

1. Introduction

The prediction of human gaze behavior is a computer vision problem with clear application to the design of more interactive systems that can anticipate a user's attention. In addition to addressing basic cognitive science questions, gaze modeling has applications in human-computer interaction (HCI) and Augmented/Virtual Reality (AR/VR) systems [2, 12, 24], non-invasive healthcare [35, 43], visual display design [18, 40], robotics [17, 25], education [3, 11, 39], etc. It is likely that gaze modeling will be ubiquitous in all HCI interfaces in the near future. Hence, there is a need for gaze prediction models that are reliable and effective yet efficient and capable of fast inference for use with edge devices (e.g., smartglasses, AR/VR headsets).

Our focus is on goal-directed behavior (not free viewing), and specifically the prediction of gaze fixations as a person searches for a given target-object category (e.g., a clock or fork). Visual search is engaged in innumerable task-oriented human activities (e.g., driving) in both real and AR/VR environments. The gaze prediction problem for visual search takes an image input and outputs a sequence of eye movements conditioned on the target. Existing ar-

chitectures for search fixation prediction use either panoptic segmentation maps [45] or object detection modules [8, 46] to encode specific search targets. However, the applicability of these methods is essentially limited to the relatively small number of target objects for which panoptic segmentation or object detector models can be realistically trained. For example, to predict the fixations in search for "pizza cutter", which is an object category not included in the dataset used to train the backbone detector in [8], a new "pizza cutter" detector would need to be trained, meaning this and related methods do not scale beyond their training data. Relatedly, existing approaches [8, 45, 46] have required the laborious collection of large-scale datasets of search behavior for each target category ($\sim$ 3000 scanpaths per category) included in the dataset, an approach that is unscalable to the innumerable potential targets that humans search for in the wild.

We therefore introduce a new problem that we refer to as *ZeroGaze*, which is an extension of zero-shot learning to the gaze prediction problem. Under ZeroGaze, a model must predict the fixations that a person makes while searching for a target (e.g., a fork) despite the unavailability of search fixations (e.g., a person looking for a fork) for model training. A more challenging version of the problem is when there are no trained detectors for the target (e.g., a fork detector) either. This is in contrast to the traditional "GazeTrain" setting (where the model has been trained on gaze behavior for all the categories that it encounters during inference). To address the ZeroGaze problem, we propose *Gazeformer*, a novel multimodal model for scanpath prediction. Gazeformer is scalable because it does not require training a backbone detector on the fixation behavior of people searching for specific object categories. We use a linguistic embedding of the target object from a language model (RoBERTa [31]) and a transformer encoder-decoder architecture [42] to model the joint image-target representations and the spatio-temporal context embedded within it. Given that recent language models can encode any object using any text string, they provide a powerful and scalable solution to the representation of target categories for which no gaze data is available for training. Additionally, Gazeformer decodes the fixation time-steps in *parallel* using a transformer decoder, providing significantly faster inference than any other method, which is necessary for gaze tracking applications in HCI products (especially in wearable edge devices). We show the advantages of our model in Fig. 1. The specific contributions of this study are:

1. We introduce the *ZeroGaze* task, where a model must predict the search fixation scanpath for a new target category without training on prior knowledge of the gaze behavior to that target.
2. We devise a novel multimodal transformer-based architecture called *Gazeformer*, which learns the interaction between the image input and the language-based semantic features of the target.
3. We propose a novel and effective scheme of fixation modeling that uses a set of Gaussian distributions in continuous image space rather than learning multinomial probability distributions over patches, resulting in an intuitive distance-based objective function.
4. We achieve a new state-of-the-art in search scanpath prediction. It outperforms baselines, not only in our new ZeroGaze setting, but also in the traditional setting where models are trained on the gaze behavior (GazeTrain) and in target-absent search. *Gazeformer* also generalizes to unknown categories while being up to more than 5 times faster than previous methods.

2. Related Work

Gaze Prediction for Search Task. Beginning with [21], most efforts for gaze prediction have been focused on using saliency maps [4, 20, 27–29] to predict free-viewing behavior. In free-viewing, attention is driven entirely by features extracted from the visual input, whereas in a search task, a top-down goal is used as well to guide attention. An early study predicted search fixations for two target categories, microwave ovens and clocks, in the identical scene contexts [47, 48]. Inverse Reinforcement Learning was used to predict search fixations in [45], exploiting a recent large-scale dataset of search fixations encompassing 18 targets categories (COCO-Search18 [9]). [8] proposed a model that predicts scanpaths using task-specific attention maps and performs well in several visual tasks including visual search. [10, 46] addressed the target-absent problem.

Zero Shot Learning. In machine learning, Zero Shot Learning refers to a problem whereby samples encountered during inference time are from classes that have not been observed during training. Zero Shot Learning has been explored in the context of image classification [16], object detection [38] and many more computer vision tasks, and Zhang *et al*. [49] extended it to zero-shot visual search by proposing a network that predicts scanpaths for unseen novel objects. However, their model requires seeing a *visual image of an exemplar of the target class* as input. In contrast, we focus on the problem of "ZeroGaze" prediction, where the model must infer scanpaths for targets without relying on template matching or fixation data specific to the target category.

Transformers. Transformers [42] use an attention mechanism to provide context for any position in the input sequence. Recently, transformers have made considerable impact in Computer Vision [15, 32, 41]. A particular line of transformer-based models (DETR [6], MDETR [23]) use a convolutional back-end to extract local visual features,

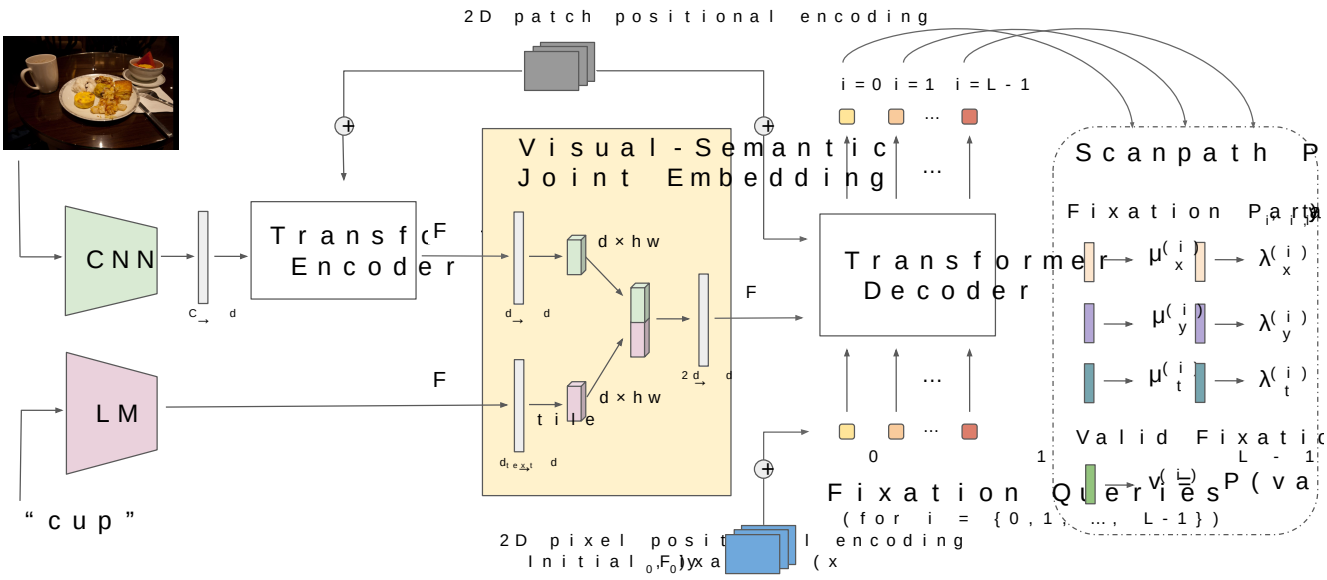

Figure 2. **Overall architecture of the Gazeformer model**. A transformer encoder block is used to contextualize CNN-extracted visual features. After being jointly embedded with target semantic features from a language model (LM), the resultant features interact with L fixation queries in a transformer decoder. The L output encodings are passed to 7 MLP layers to obtain coordinates, duration and validity for all possible fixations in parallel. We use ResNet-50 [19] as the CNN and RoBERTa [31] as the LM for our experiments.

which are then used in a transformer encoder-decoder architecture to perform contextual reasoning. A recent work [7] predicts free-viewing viewport scanpath for 360° videos with a transformer encoder based architecture. To our best knowledge, ours is the first study that applies a *transformer encoder-decoder* architecture with *multimodal understanding* to predict human scanpaths for *visual search*.

3. Gazeformer

Two factors enable Gazeformer to generalize to the ZeroGaze scanpath prediction problem. First, Gazeformer jointly embeds visual and semantic features of a search target, with the latter extracted from a language model (RoBERTa [31]). This means that the target is encoded as an interaction between image features and real-world semantic relationships. Second, the self-attention mechanism of the encoder allows it to learn the global and local context of the scene that previous convolution-based methods missed. Gazeformer learns where targets are usually located, meaning that it learns object co-occurrence and other object-scene relationships that convey information about target location (e.g., first finding a kitchen counter top when searching for a microwave), and this enables it to generalize well to target-absent search, where scene context is crucial. Two additional factors improve Gazeformer's scanpath prediction capabilities. First, our model uses a transformer-based decoder to predict an entire sequence of search fixations in parallel (including when to terminate). This parallel decoding not only offers a considerable speed up in inference time for scanpath prediction, but also allows the model to learn long-range dependencies between fixations in a bidirectional manner, differing from the standard unidirectional sequential scanpath prediction approach [8,45,46]. Second, instead of performing a patch-wise prediction for generating a sequence of fixations, as in prior approaches [8,45,46], we directly regress the fixation locations parameterized by Gaussian distributions. This enables Gazeformer to learn an effective distance-based objective function for predicting the entire sequence of scanpath fixations at once.

3.1. Architecture

The overview of Gazeformer architecture is provided in Fig. 2. First, image features are extracted from a ResNet-50 backbone and further processed through a transformer encoder to obtain contextual image features $\mathbf{F}_{image}$. The semantic embedding $\mathbf{F}_{target}$ of a search target is extracted from RoBERTa. Consequently, we compute $\mathbf{F}_{joint}$, a joint visual-semantic embedding of image and target features. Given the features $\mathbf{F}_{joint}$ and learnable fixation queries (which represent the time step information of each fixation), a transformer decoder produces a sequence of fixation embeddings *in parallel* for each time-step. The fixation embeddings are then processed through seven independent Multi-Layer Perceptron (MLP) layers, which yield (1) fixation coordinates, (2) fixation duration, and (3) scanpath termination information as final model output. Below are the details of each model component.

Image Feature Encoding. We extract features for an image

by resizing it to 1024×640 resolution and passing it through a ResNet-50 [19] network to extract a feature map of dimensions $C\times h\times w$, where $C = 2048, h = 20, w = 32$. We intentionally choose the resize resolution to be 1024×640 so that $h = 20$ and $w = 32$, in order to have the same input granularity as baseline representations [45, 46]. We flatten the spatial dimensions of this feature map to obtain a 2D feature tensor of size $C\times hw$. Since $C = 2048$ is prohibitively large for further computation, we use a linear layer to reduce the number of feature dimensions from 2048 to a smaller value d. This forms the input to a transformer encoder block that consists of a cascade of N_{enc} standard transformer encoder [42] layers. Each layer uses multi-head self-attention, feed-forward networks, and layer normalization to find contextual embeddings of the input. To indicate the location of each patch, we use a fixed sinusoidal $2D$ positional encoding as in [6]. This encoder block creates a task-agnostic contextualized representation of the image in the form of feature tensor $\mathbf{F}_{image} \in \mathbb{R}^{d\times hw}$.

Target Feature Semantic Encoding. To extract target features, we use the language model RoBERTa [31] (we use the RoBERTa-base variant) to encode the text string signifying the target category (such as "potted plant") as a tensor $\mathbf{F}_{target} \in \mathbb{R}^{d_{text}}$, where $d_{text} = 768$. Unlike word embedding frameworks like Word2vec [33] and GloVe [36], RoBERTa can encode both single-word category names such as "car" and multi-word target category names such as "stop sign" to a fixed length vector of dimension d_{text}.

Joint Embedding of Image and Target Feature. To create a joint visual-semantic embedding space of image and target features, we first independently map $\mathbf{F}_{image}$ and $\mathbf{F}_{target}$ to a shared multimodal d-dimensional latent space using modality-specific linear transformations. Then we tile the modality-specific transformation of $\mathbf{F}_{target}$ spatially hw times. We concatenate the transformed image features and the transformed and tiled target features (both now sized $d\times hw$) along the channel dimension and obtain a 2D tensor of size $2d\times hw$, which is again linearly projected (with ReLU activation) to the dimension size of d. The final jointly embedded visual-semantic features are $\mathbf{F}_{joint} \in \mathbb{R}^{d\times hw}$. This joint embedding step differs from previous multimodal methods [23, 44] in that we append semantic features of a search target only after a target-agnostic contextual image representation is obtained from the encoder.

Fixation Decoding. We decode an entire sequence of fixations in *one go* using a transformer decoder. Like most transformer architectures, we have a maximum output scanpath sequence length L which requires fixation sequences to be *padded* if their lengths are smaller than L. The decoder contains N_{dec} standard transformer decoder layers [42] (stacked self-attention and cross-attention layers with minor modifications) and processes $\mathbf{F}_{joint}$ along with a set of L *fixation queries* $Q_i \in \mathbb{R}^d, i \in \{0, 1, ..., L-1\}$. The fixation queries are randomly initialized learnable embeddings that provide fixation time step information. At each decoder layer, the latent fixation embeddings interact with each other through self-attention, and also interact with $\mathbf{F}_{joint}$ through encoder-decoder cross-attention. Note that before each cross-attention step, a fixed 2D positional encoding is added to $\mathbf{F}_{joint}$ in order to provide positional information about the patches. We encode initial fixation location (x_0, y_0) (which affects the generated scanpath) using another fixed 2D positional encoding and add it to Q_0. The output of the transformer decoder block is $\mathbf{F}_{dec} \in \mathbb{R}^{d\times L}$.

Scanpath Prediction. Previous scanpath prediction models [8, 45, 46] predict fixations on a patch-level granularity by reducing the image space into a set of discrete locations and generating a multinomial probability distribution over them. However, one limitation of this approach is that all patches are at the same distance from each other—patches closer to each other are treated the same way as patches that are further apart. To remedy this, we propose to directly *regress* the raw fixation co-ordinates for each L possible time steps from $\mathbf{F}_{dec}$. To incorporate the inter-subject variability in human fixations, we model fixation locations and durations using Gaussian distributions, and consequently regress the mean and log-variance of 2D co-ordinates and duration of a fixation using six separate MLP layers. For the i^{th} fixation, let $\mu_{x_i}, \mu_{y_i}, \mu_{t_i}, \lambda_{x_i}, \lambda_{y_i}, \lambda_{t_i}$ denote the mean and log-variance for the x and y positions and duration t. Using the reparameterization trick [26], the fixation co-ordinates x_i and y_i and duration t_i are estimated as follows:

$$x_i = \mu_{x_i} + \epsilon_{x_i} \cdot \exp(0.5\lambda_{x_i}), \quad y_i = \mu_{y_i} + \epsilon_{y_i} \cdot \exp(0.5\lambda_{y_i}),$$
$$t_i = \mu_{t_i} + \epsilon_{t_i} \cdot \exp(0.5\lambda_{t_i}), \quad \epsilon_{x_i}, \epsilon_{y_i}, \epsilon_{t_i} \in \mathcal{N}(0, 1). \quad (1)$$

This allows our network to be fully differentiable despite having a probabilistic component. Finally, since we use padding, we make a prediction for each output if the current step belongs to a valid fixation or a padding token. This is done by an MLP classifier with softmax activation which classifies each of the L slices of $\mathbf{F}_{dec}$ to be a valid fixation or padding token. During inference, we collect the sequence of (x_i, y_i, t_i, v_i) ordered quads for $i \in \{0, 1, ..., L-1\}$, where x_i, y_i are the coordinates of fixation i, t_i is the duration of fixation i, and v_i the probability of being a valid fixation. We traverse the sequence from $i = 0$ to $i = L-1$ and terminate the sequence at the first padding token ($v_i < 0.5$). We experimentally confirmed that our model performs better with regression than with classification. Architecture details are in the supplemental.

3.2. Training

We view scanpath prediction as a sequence modeling task. The total multitask loss $\mathcal{L}$ to train our network with backpropagation for a minibatch of M samples is:

$$\mathcal{L} = \frac{1}{M}\sum_{j=1}^{M}\left(\mathcal{L}_{xyt}^{j} + \mathcal{L}_{val}^{j}\right), \tag{2}$$

$$\text{where } \mathcal{L}_{xyt}^{j} = \frac{1}{l^j}\sum_{i=0}^{l^j-1}\left(|x_i^j - \hat{x}_i^j| + |y_i^j - \hat{y}_i^j| + |t_i^j - \hat{t}_i^j|\right),$$

$$\mathcal{L}_{val}^{j} = -\frac{1}{L}\sum_{i=0}^{L-1}\left(v_i^j \log \hat{v}_i^j + (1 - v_i^j)\log(1 - \hat{v}_i^j)\right).$$

Here $s^j = \{(x_i^j, y_i^j, t_i^j)\}_{i=0}^{L-1}$ is the predicted scanpath and L is the maximum scanpath length. l^j is the length of the ground truth scanpath $\hat{s}^j = \{(\hat{x}_i^j, \hat{y}_i^j, \hat{t}_i^j)\}_{i=0}^{l^j-1}$. $\hat{v}_i^j$ is a binary scalar representing ground truth of i^{th} token in s^j being a valid fixation or padding and v_i^j is the probability of that token being a valid fixation as estimated by our model. The losses included in $\mathcal{L}_{xyt}$ are the L_1 regression losses for x and y co-ordinates and duration t of the fixations. $\mathcal{L}_{val}$ is the negative log likelihood loss for validity prediction for each token. We find the L_1 loss to be more appropriate for scanpath prediction because of the intrinsic variability in human fixation locations in a multi-subject setting. Note that we mask out ground truth padded tokens while calculating $\mathcal{L}_{xyt}$ to only account for valid fixations in ground truth. Optimization and training details are in the supplemental.

4. Experiments

We evaluate and compare the model performance under two prediction scenarios: ZeroGaze and GazeTrain. In the ZeroGaze setting, a model must predict search fixations for a category not encountered during training. GazeTrain is the traditional scanpath prediction setting where a model is trained and tested to predict search fixations using the same set of target categories. All experiments pertaining to ZeroGaze and GazeTrain settings were conducted using the target-present data of the COCO-Search18 [9] dataset following the evaluation scheme in [45]. COCO-Search18 consists of 3101 target-present (TP) images, and 100,232 scanpaths collected for 18 categories from 10 subjects. We use the original train-validation-test split provided by [45] and report all results on the test set. We also demonstrate the superior generalizability of our method by comparing it with other baselines on the target-absent data of COCO-Search18 after training the model on either (1) the target-present data or (2) the target-absent data. The IRL model from Yang *et al*. [45], the model from Chen *et al*. [8] and a recent FFM model from Yang *et al*. [46] are used as baseline methods. Since only Chen *et al*. [8]'s model predict duration, we also provide a variation of Gazeformer called *Gazeformer-noDur* which does not train on or predict fixation duration for fair comparison with the IRL and FFM baselines. We remap the predicted fixations from Chen *et al*. [8]'s model to a 20×32 grid following [46]. All of the models except IRL predict search stopping behavior, and we use the baselines' stopping prediction modules whenever available to obtain a fairer comparison. We try to maintain the same hyperparameters across settings for all models to evaluate generalizability and scalability.

4.1. Implementation Details

Unless explicitly specified otherwise, the results reported in this paper are for the Gazeformer model with 6 encoder layers, 6 decoder layers, 8 attention heads per multihead attention block, and hidden size (d) of 512. Following [45], we set maximum scanpath length to 7 (including the initial fixation) for our experiments. Due to the limited availability of training data and computational resources for additionally finetuning the backbone networks, we use frozen ResNet-50 and RoBERTa encodings. Models generate 10 scanpaths per test image by sampling at each fixation, and the reported metrics are averages.

4.2. Metrics

We report model performance in search scanpath prediction using a diverse set of metrics. *Sequence Score (SS)* [45] converts scanpaths into strings of fixation cluster IDs and a string matching algorithm [34] measures similarity between two strings. *Semantic Sequence Score (SemSS)* [46] modifies SS by converting scanpaths to strings of fixated objects in the scene instead of cluster IDs, increasing interpretability. Unlike the implementation of SemSS by [46] that disregarded the COCO stuff classes, we include both thing and stuff classes for object assignment. We also include *Fixation Edit Distance (FED)* and *Semantic Fixation Edit Distance (SemFED)*, which convert scanpaths to strings like the SS and SemSS metrics, respectively, but use the Levenshtein algorithm [30] to measure scanpath dissimilarity. *Multimatch (MM)* [1, 14] is a popular metric that measures the scanpath similarity at the pixel level. Here, MM indicates the average of the shape, direction, length, and position scores (individual metrics in supplemental). *Correlation Coefficient (CC) [22]* measures the correlation between the normalized model and the human fixation map (i.e., the 2D distribution map of all fixations convolved with a Gaussian). *Normalized Scanpath Saliency (NSS) [37]* is a discrete approximation of CC, which averages the values of a model's fixation map from the human fixated locations [5]. Higher values for SS, SemSS, MM, NSS and CC indicate greater similarity between model-generated and human search scanpaths. The direction is opposite for FED and SemFED. We include SS, FED, SemSS and SemFED both with duration (as in [13]) and without duration.

	SS↑		SemSS↑		FED↓		SemFED ↓		MM	CC	NSS
	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	↑	↑	↑
IRL [45]	0.290	-	0.314	-	4.606	-	4.377	-	0.774	0.241	4.018
Chen *et al.* [8]	0.210	0.041	0.211	0.034	5.720	210.498	5.608	211.636	0.717	0.002	0.001
FFM [46]	0.300	-	0.334	-	3.271	-	2.918	-	0.731	0.271	**5.247**
Gazeformer-noDur	**0.359**	-	**0.391**	-	2.788	-	2.474	-	**0.822**	0.316	4.671
Gazeformer	0.358	**0.312**	**0.391**	**0.348**	**2.766**	**12.505**	**2.438**	**10.391**	0.812	**0.324**	4.929

(a)

	SS↑		SemSS↑		FED↓		SemFED ↓		MM	CC	NSS
	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	w/o Dur	w/ Dur	↑	↑	↑
Human	0.490	0.409	0.548	0.456	2.531	11.526	1.637	8.086	0.857	0.472	8.129
IRL [45]	0.418	-	0.499	-	2.722	-	2.182	-	0.833	0.434	6.895
Chen *et al.* [8]	0.451	0.403	0.504	0.446	2.187	10.795	1.788	8.782	0.820	0.547	6.901
FFM [46]	0.392	-	0.443	-	2.693	-	2.284	-	0.808	0.370	5.576
Gazeformer-noDur	**0.504**	-	**0.534**	-	**2.061**	-	**1.742**	-	0.849	0.559	8.356
Gazeformer	**0.504**	**0.451**	0.525	**0.485**	2.072	**9.708**	1.810	**7.688**	**0.852**	**0.561**	**8.375**

(b)

Table 1. Model performance comparison under (a) ZeroGaze setting, (b) traditional GazeTrain setting. Best performance is highlighted in bold. Performances that exceed human consistency are underlined.

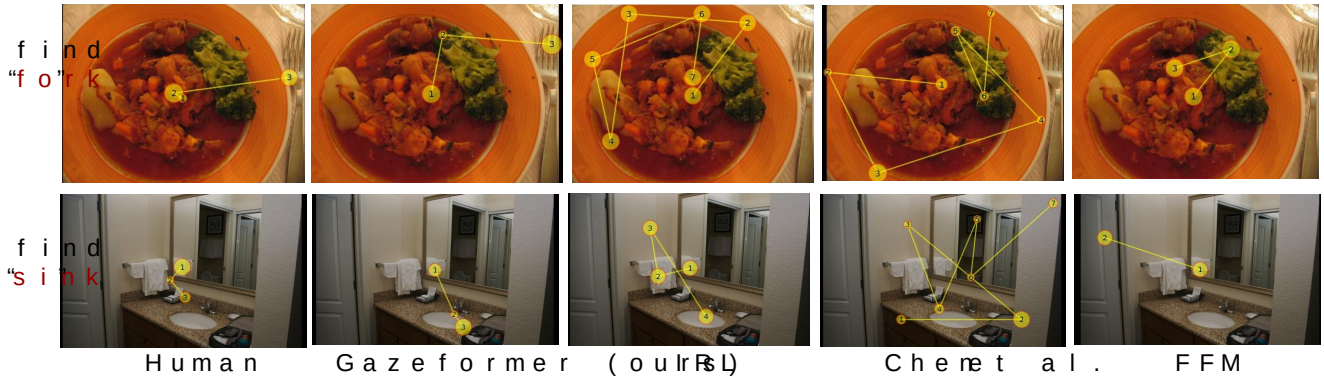

Figure 3. Visualization of scanpath prediction under the ZeroGaze setting. The number and radius indicate the fixation order and duration (if predicted), respectively. Our Gazeformer model predicts efficient, human-like scanpaths.

4.3. ZeroGaze Setting

Models were trained on 17 of the COCO-Search18 categories and tested on the one left-out category in a cross-validation manner. For example, to evaluate ZeroGaze performance on category $\mathcal{C}$, we remove the category $\mathcal{C}$ scanpaths from the training data and predict gaze behavior for this category during testing. Thus, we ensure that a model does not see any behavioral data related to category $\mathcal{C}$ for ZeroGaze prediction. Performance was averaged over all categories and weighted by the number of test cases for each category. As shown in Table 1(a), Gazeformer achieves state-of-the-art ZeroGaze performance, outperforming the baselines by considerable margins across almost all metrics. See Fig. 3 for some qualitative results. More qualitative results can be found in the supplemental.

4.4. Traditional GazeTrain Setting

In the traditional GazeTrain setting, models were trained with target-present search fixations on 18 different target categories (from the COCO-Search18 dataset) and tested on the same set of categories but with unseen test cases. We compared the predictive performance of Gazeformer and previous models of search scanpath prediction. Table 1(b) shows that Gazeformer generally outperforms the baselines (by significant margins for some metrics) and is at (or slightly exceeds) a noise ceiling imposed by human consistency rates (i.e., predictions cannot be much better). Hence, Gazeformer is the new state-of-the-art in target-present visual search scanpath prediction. We include qualitative results of Gazeformer and other baseline models in the supplemental.

	Trained on Target-Present					Trained on Target-Absent				
	SS↑		SemSS↑		MM	SS↑		SemSS↑		MM
	w/o Dur	w/ Dur	w/o Dur	w/ Dur	↑	w/o Dur	w/ Dur	w/o Dur	w/ Dur	↑
Human	0.398	0.369	0.436	0.404	0.838	0.398	0.369	0.436	0.404	0.838
IRL [45]	0.304	-	0.349	-	0.808	0.323	-	0.378	-	0.805
Chen *et al.* [8]	0.350	0.330	0.395	0.380	0.813	0.345	0.323	0.347	0.335	0.799
FFM [46]	0.360	-	0.413	-	0.814	0.362	-	0.413	-	0.814
Gazeformer-noDur	0.366	-	**0.419**	-	**0.833**	0.369	-	0.422	-	0.830
Gazeformer	**0.368**	**0.356**	**0.419**	**0.399**	0.825	**0.375**	**0.361**	**0.438**	**0.417**	**0.844**

Table 2. Performance comparison for models tested on target-absent data after training on target-present data and target-absent data. Best performance is highlighted in bold. Performances that exceed human consistency are underlined.

4.5. Generalizability to Target-Absent Search

To demonstrate generalizability, we evaluated model performance on COCO-Search18 target-absent trials under two settings - (1) where the model is trained on only target-present data, and (2) where the model is trained on only target-absent data. The first setting should be direct evidence for the capability of the model to learn the important contextual relationships between scene objects that humans use to guide their search for a target when it is absent from the scene. The second setting tests the model's capability to learn search behavior for the target-absent task by training on target-absent human trials. We report the results for both settings in Table 2. Similar to the target-present results, Gazeformer again outperforms all baselines in multiple metrics, establishing new SOTA in target-absent scanpath prediction as well (although the human noise ceiling was only achieved with training on the target-absent data). Gazeformer appears to be learning scene context and object relationships well and generalizes to target-absent search without any change in architecture or hyperparameters. We include a comprehensive version of Table 2 and qualitative evidence for Gazeformer's use of context through attention maps in the supplemental.

4.6. Inference Speed

	Time (in ms)↓	Inferences/s ↑	Speedup ↑
Chen *et al.* [8]	386	2.59	1X
FFM [46]	133	7.52	2.9X
IRL [45]	85	11.77	4.5X
Gazeformer	**68**	**14.71**	**5.7X**

Table 3. Comparison of inference speeds. Gazeformer achieves 5.7X speedup over highly competent Chen *et al.*'s [8] model.

Despite clearly outperforming baselines in search scanpath prediction, Gazeformer is many times faster. We attribute this to the parallel decoding mechanism in the transformer decoder that predicts the entire scanpath in one go while other baselines are sequential in nature. We report the average time measured on COCO-Search18's test split of target-present trials in Table 3. We measure time in the real-world setting where the model receives and operates on one search task case (image-target pair) at a time.

4.7. Extension to Unknown Categories

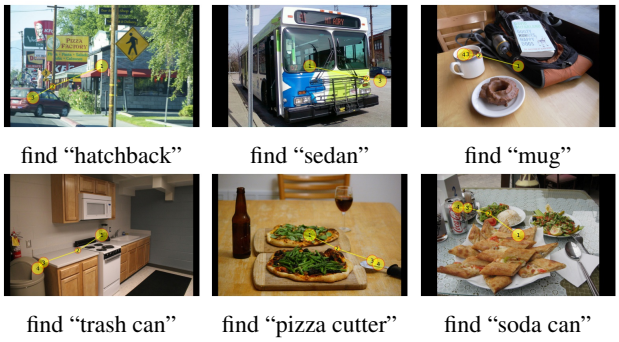

Figure 4. Generalization of Gazeformer to unknown categories. Top row shows extensions to non-canonical names of COCO-Search18's categories; bottom row shows extensions beyond COCO-annotated categories.

Since the proposed model uses the RoBERTa language model to encode target objects, hypothetically it can generalize to searching for any object that can be described in words or text. First, we generate and visualize new scanpaths using the synonyms or hyponyms of the COCO-Search 18 objects to define targets (e.g., replacing "cup" with "mug" and "car" with "hatchback" or "sedan"). We also investigate if our model can extend beyond the MS-COCO dataset used to train backbone models, to completely unseen targets like "trash can", "pizza cutter" or "soda can". Quantitative comparison is not possible because there are no human search fixation data, so we only generate and visualize scanpaths for these cases. As shown in Fig. 4, Gazeformer generates plausible natural-looking scanpaths that successfully find the unknown target. Previ-

ous methods are confined to a predefined set of categories due to their reliance on detector/panoptic maps and fail to handle these scenarios.

4.8. Impact of Language Embeddings

Gazeformer uses RoBERTa [31] linguistic embeddings as target representations, and to explore their role in model performance, we replaced the 18 target category embeddings from RoBERTa with fixed random embeddings.

Target Embedding	SS↑	FED↓	NSS↑
RoBERTa	**0.359**	**2.788**	**4.671**
Fixed Random	0.336	2.873	4.524

(a)

Category	Embedding	SS↑	FED↓	NSS↑
stop sign	RoBERTa	0.430	2.256	4.559
	Random	0.358	2.176	4.258
clock	RoBERTa	0.399	2.600	2.110
	Random	0.310	3.274	1.168
cup	RoBERTa	0.354	2.942	1.483
	Random	0.357	2.724	1.547

(b)

Table 4. (a) ZeroGaze results for RoBERTa and fixed random embeddings. (b) Comparison of ZeroGaze results within categories.

Figure 5. Gazeformer often confuses a target with semantically similar objects under the ZeroGaze setting, e.g., being distracted by a bottle or bowl when the new target category is a cup.

Table 4(a) shows that RoBERTa embeddings improve Gazeformer's performance under the ZeroGaze setting. Interestingly, Table 4(b) shows that language embeddings for different targets produce different benefits, with "stop sign" benefiting the most and "cup" the least. We speculate that this is due to how target words used in natural language differ in their semantic consistency. For example, the word "cup" is used in various contexts, often interchangeably with multiple different synonyms, e.g., drink, bowl, container. Indeed, we find that the model is often confused with other semantically similar objects when searching for a target; this is especially true when the target is a cup (Fig. 5). Note that even with fixed random embeddings, Gazeformer achieves slightly better performance with respect to the FFM and IRL models. We posit that this is due to the ability of the Gazeformer transformer encoder block to explicitly capture scene context which enables scene exploration - a considerable part of the search process. Nevertheless, semantic cues from the language model embeddings boost the performance of our model even further.

4.9. Pixel Regression vs Grid Classification

In Table 5, we demonstrate the impact of regressing the fixation co-ordinate parameters as compared to the more standard method of classifying amongst patches. We replace the regression step of Gazeformer with a classifier MLP that learns a probability distribution over the 20×32 patches (henceforth, called *Gazeformer-noReg*).

	SS↑	FED↓	NSS↑
Gazeformer	**0.504**	**2.072**	**8.375**
Gazeformer-noReg	0.477	2.158	7.545

Table 5. Gazeformer performance compared to a model variant where pixel regression is replaced by patch classification (*Gazeformer-noReg*). Results are from GazeTrain setting.

Table 5 shows that directly regressing the fixation locations provides significantly better performance, as hypothesized. Also note that even when the model estimates a grid probability distribution (similar to previous methods), the performance is still superior to baselines. This highlights the strength of our core architecture.

5. Conclusion

We introduced *ZeroGaze*, a novel attention task aimed at scalability whereby a model must predict the fixation scanpaths for search targets despite having no prior search fixations available for training. We proposed a new multimodal, transformer-based model called *Gazeformer* coupled with a novel scanpath prediction method, which not only showed impressive ZeroGaze results but also scaled well to uncommon and unknown target categories. Our model also achieved new state-of-the-art scanpath prediction performance in traditional target-present and target-absent search, while having faster inference speeds. Because Gazeformer can generate fixation locations and durations for an entire scanpath in negligible time, it enables anticipation of *where* and *when* a person's attention will shift, thus making it ideal for time-critical HCI applications. We expect that Gazeformer's scalability, effectiveness, and speed will fuel the use of gaze prediction models in HCI applications and products. In future work, we look forward to extensions of Gazeformer to other visual tasks like free-viewing and VQA.

Acknowledgement. This project was supported by US National Science Foundation Awards IIS-1763981, IIS-2123920, NSDF DUE-2055406, and the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.

References

[1] Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior research methods*, 47(4):1377–1392, 2015. 5

[2] Christopher R Bennett, Peter J Bex, and Lotfi B Merabet. Assessing visual search performance using a novel dynamic naturalistic scene. *Journal of Vision*, 21(1):5–5, 2021. 1

[3] Maël Beuget, Sylvain Castagnos, Christophe Luxembourger, and Anne Boyer. Eye gaze sequence analysis to model memory in e-education. In *International Conference on Artificial Intelligence in Education*, 2019. 1

[4] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013. 2

[5] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 5

[6] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*. Springer, 2020. 2, 4

[7] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021. 3

[8] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3, 4, 5, 6, 7

[9] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021. 2, 5

[10] Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *Proceedings of CVPR International Workshop on Gaze Estimation and Prediction in the Wild*, 2022. 2

[11] Neila Chettaoui, Ayman Atia, and Med Salim Bouhlel. Student performance prediction with eye-gaze data in embodied educational context. *Education and Information Technologies*, 28(1):833–855, 2023. 1

[12] Jiyong Chung, Hyeokmin Lee, Hosang Moon, and Eunghyuk Lee. The static and dynamic analyses of drivers' gaze movement using vr driving simulator. *Applied Sciences*, 12(5):2362, 2022. 1

[13] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010. 5

[14] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012. 5

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[16] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 2

[17] Kenko Fujii, Gauthier Gras, Antonino Salerno, and Guang-Zhong Yang. Gaze gesture based human robot interaction for laparoscopic surgery. *Medical image analysis*, 44:196–214, 2018. 1

[18] Tim Halverson and Anthony J Hornof. A minimal model for predicting visual search in human-computer interaction. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, 2007. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3, 4

[20] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015. 2

[21] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. 2

[22] Timothée Jost, Nabil Ouerhani, Roman Von Wartburg, René Müri, and Heinz Hügli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005. 5

[23] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2, 4

[24] Sebastian Kapp, Michael Barz, Sergey Mukhametov, Daniel Sonntag, and Jochen Kuhn. Arett: augmented reality eye tracking toolkit for head mounted displays. *Sensors*, 21(6):2234, 2021. 1

[25] Heecheol Kim, Yoshiyuki Ohmura, and Yasuo Kuniyoshi. Gaze-based dual resolution deep imitation learning for high-precision dexterous robot manipulation. *IEEE Robotics and Automation Letters*, 6(2):1630–1637, 2021. 1

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 4

[27] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 2

[28] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps

trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 2

[29] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 2

[30] Vladimir I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics. Doklady*, 10:707–710, 1965. 5

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3, 4, 8

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 2

[33] Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013. 4

[34] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 5

[35] Domen Novak and Robert Riener. Enhancing patient freedom in rehabilitation robotics using gaze-based intention detection. In *2013 IEEE 13th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2013. 1

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014. 4

[37] Robert J Peters, Asha Iyer, Christof Koch, and Laurent Itti. Components of bottom-up gaze allocation in natural scenes. *Journal of Vision*, 5(8):692–692, 2005. 5

[38] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Asian Conference on Computer Vision*. Springer, 2018. 2

[39] Sanjay Rebello, Minh Hoai, Yang Wang, Tianlong Zu, John Hutson, and Lester Loschky. Machine learning predicts responses to conceptual questions using eye movements. In *Proceedings of the Physics Education Research Conference*, 2018. 1

[40] Shigeo Takahashi, Akane Uchita, Kazuho Watanabe, and Masatoshi Arikawa. Gaze-driven placement of items for proactive visual exploration. *Journal of Visualization*, 25(3):613–633, 2022. 1

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 2021. 2

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, **L** ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2, 4

[43] Mélodie Vidal, Jayson Turner, Andreas Bulling, and Hans Gellersen. Wearable eye tracking for mental health monitoring. *Computer Communications*, 35(11):1306–1311, 2012. 1

[44] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[45] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2, 3, 4, 5, 6, 7

[46] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *European Conference on Computer Vision*, 2022. 2, 3, 4, 5, 6, 7

[47] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[48] Gregory J Zelinsky, Yupei Chen, Seoyoung Ahn, Hossein Adeli, Zhibo Yang, Lihan Huang, Dimitrios Samaras, and Minh Hoai. Predicting goal-directed attention control using inverse-reinforcement learning. *Neurons, behavior, data analysis and theory*, 2020. 2

[49] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any waldo with zero-shot invariant and efficient visual search. *Nature communications*, 9(1):1–15, 2018. 2