Look Hear: Gaze Prediction for Speech-directed Human Attention

Sounak Mondal[1], Seoyoung Ahn[2], Zhibo Yang[3], Niranjan Balasubramanian[1], Dimitris Samaras[1], Gregory Zelinsky[1], and Minh Hoai[4]

[1] Stony Brook University, NY, USA
[2] UC Berkeley, CA, USA
[3] Waymo LLC
[4] The University of Adelaide, Adelaide, Australia

Abstract. For computer systems to effectively interact with humans using spoken language, they need to understand how the words being generated affect the users' moment-by-moment attention. Our study focuses on the incremental prediction of attention as a person is seeing an image and hearing a referring expression defining the object in the scene that should be fixated by gaze. To predict the gaze scanpaths in this *incremental object referral* task, we developed the *Attention in Referral Transformer* model or *ART*, which predicts the human fixations spurred by each word in a referring expression. ART uses a multimodal transformer encoder to jointly learn gaze behavior and its underlying grounding tasks, and an autoregressive transformer decoder to predict, for each word, a variable number of fixations based on fixation history. To train ART, we created *RefCOCO-Gaze*, a large-scale dataset of 19,738 human gaze scanpaths, corresponding to 2,094 unique image-expression pairs, from 220 participants performing our referral task. In our quantitative and qualitative analyses, ART not only outperforms existing methods in scanpath prediction, but also appears to capture several human attention patterns, such as waiting, scanning, and verification. Code and dataset are available at: `https://github.com/cvlab-stonybrook/ART`.

Keywords: Scanpath Prediction · Object Referral · Human Attention

1 Introduction

Humans are unique in that we use language to direct each others' attention in visual tasks. For example, a customer telling a baker "I'd like the smallest pastry on the left" communicates the desired object that needs to be selected. Understanding the human capacity to use these *referring expressions* to incrementally direct attention is an important problem in cognitive science and has been studied for over half a century, with the more recent studies adopting eye-tracking methods [3,16,32,58]. Most relevant is work showing the very tight link between a word in a referring expression and the very next eye movements of the person hearing it [29,59], suggesting that humans *incrementally* integrate visual information and word-by-word linguistic guidance in our attention control.

However, these studies are limited in that they used small numbers of simple objects (often line drawings) in arrays (not scenes) and this constrained the linguistic complexity of the referring expression. How spoken language guides another person's attention in more naturalistic and ecologically valid contexts is still an open question in cognitive science.

As our interactions with computers, vehicles, and AR/VR devices deepen, human-computer interaction (HCI) systems also need to give spoken guidance to users that is similarly effective in directing their attention. But, to attain this degree of synchrony with users, HCI systems must be able to integrate vision and language inputs to predict human gaze. Applications with this predictive ability will be highly time-sensitive and crucial for activities such as voice-assisted VR driving, offering a streamlined and immersive user experience where a person's gaze can be directed by generated spoken language as if the interaction were with another person. Being able to incrementally predict how a user integrates their visual input with a spoken instruction to direct their attention is a general advance in speech-assisted HCI, benefiting a broad range of applications, including efficient foveated rendering [52], VR sickness reduction [1], AR/VR eye–hand coordination analysis during human-object interaction [36], VR skill training/assessment (*e.g.*, driving [35], surgery [5]), and user engagement analysis [31]. Using predicted gaze for each word as guidance will enable speech-assisted HCI systems to incrementally generate efficient and clear instructions, correctly guiding user attention. Measuring gaze using eye-trackers instead of predicting it is more accurate, but also costlier and has limited applicability to aforementioned scenarios (such as time-crucial HCI) which require gaze prediction, and to situations where eye-trackers are unavailable or prohibited.

In this context, we study the *incremental object referral task*, for which we *incrementally* predict eye movements of humans searching for a target object in an image as they are hearing a *referring expression* describing that target. This task has not been studied previously in the context of human gaze prediction. The standard object referral task [45, 74] requires localizing the target object given an image and a referring expression. Our task is different in that we aim to incrementally predict the attention of a human as they are hearing the expression. Our task is also related to the categorical search task [77], where humans direct their attention to a category of target object in an image. However, our incremental object referral task differs from categorical search in two key aspects. First, in incremental object referral, the target is designated by a complex referring expression (*e.g.*, "red baseball glove on the desk") since the image may contain other objects belonging to the same category as the target. Referring expressions often refer to a target by its attributes ("red") or its spatial relationships to other objects ("on the desk"), making it even more challenging to precisely localize the target without the broader descriptive context. In categorical search, the target is designated by only its category name (*e.g.*, "baseball glove"), which excludes the spatial terms and attributes commonly used by humans to describe an object. This task is also less realistic in that images depicting multiple instances of the target category are excluded (as spatial terms and attributes are not used).

Second, because categorical search studies of attention use only an one or two word category name to designate a target (*e.g.*, "baseball glove"), most of the human search fixations occur only *after* the complete referring "expression" is provided. This makes it an impoverished example of the longer and more natural referring expressions that we hope to study, ones requiring an incremental allocation of attention. Our more ecologically valid incremental object referral task therefore contributes to this cognitive science question by enabling exploration of natural referential expressions in real-world image contexts and generating testable hypotheses about how humans integrate language and vision.

Given its differences from related tasks that were studied previously, the incremental object referral task presents several technical challenges that necessitate more than mere updates to existing models. For instance, a standard object referral model cannot be re-purposed for incremental prediction. Although it is possible to sequentially input incomplete referring expressions, one word at a time, and use the predicted positions as proxies for fixation locations, this approach proves ineffective as existing object referral models train on complete referring expressions in contrast to how humans integrate visual and linguistic information in a word-by-word basis [29, 59]. Alternatively, one can adapt an existing scanpath prediction model for our incremental object referral task, but this approach is also unsuitable because existing scanpath prediction models do not learn the object grounding processes (for *both* partial and complete referring expressions) hypothesized to underlie the gaze behavior observed in our task.

To address the aforementioned challenges of the incremental object referral task, we introduce the *Attention in Referral Transformer (ART)* model. ART is tailored to the multimodal demands of our task as it uses a multimodal transformer encoder that jointly learns gaze prediction and object grounding objectives. Furthermore, we integrate an autoregressive transformer decoder that leverages fixation history to better predict the subsequent fixations corresponding to each sequentially presented word from the referring expression. This innovative decoder component flexibly adjusts both the count and the parameters of predicted fixations in alignment with the evolving input, thus mirroring the dynamic nature of human attention.

To train ART, we collected *RefCOCO-Gaze* (Fig. 1), a large-scale dataset of gaze behavior from 220 people performing the incremental object referral task on 2,094 images and associated referring expressions from RefCOCO [74] dataset. Compared to baselines [11, 50, 67], only ART was able to accurately predict the dynamic changes in gaze of humans incrementally hearing and shifting their attention in response to the words in the referring expression, even with incomplete target descriptions. Ablation studies showed that pre-training and training on auxiliary grounding tasks, such as object localization and target category prediction, improves ART's gaze prediction performance. In qualitative analyses, ART was shown to capture several fixation patterns of people performing the incremental object referral task, such as waiting, scanning, and verification, suggesting that it learned to strategically disambiguate vision-language ambiguities.

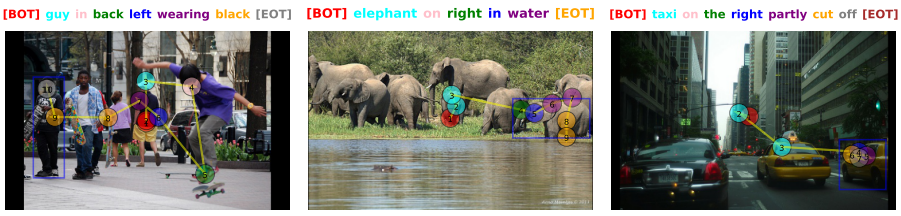

Fig. 1: RefCOCO-Gaze Dataset. Sample image-expression pairs and corresponding scanpaths under our *incremental object referral* task. Fixations (denoted by circles numbered with fixation order) are color-coded to the corresponding word in the referring expression (above each image). Fixations color-coded to `[BOT]` occurred before the expression started, and fixations color-coded to `[EOT]` occurred after the expression ended. Blue bounding boxes indicating referred objects were not visible during trials.

In summary, our contributions are: **(1)** Introducing the *incremental object referral* task for gaze prediction that will lead to more user-responsive HCI systems. **(2)** Creating *RefCOCO-Gaze*, a large-scale dataset of gaze behavior during the incremental object referral task. **(3)** Developing *ART*, the first gaze prediction model of incremental object referral that offers computational solutions to the incremental and multimodal aspects of our task. **(4)** Bringing RefCOCO-Gaze and ART into the toolboxes of researchers studying incremental object referral, thereby enabling them to understand how humans dynamically merge their visual and linguistic information in the real world to control their attention.

2 Related Work

Interest in gaze prediction as a computer vision problem has been growing [11,50, 70,71], given that the anticipation of user attention would enable more natural augmented/virtual reality systems [6,15,49,52]. Most existing human attention prediction models predict free-viewing behavior [7,25,46], but fail to generalize to goal-directed behaviors, such as visual search [22,33]. More related is the work predicting eye fixations during the search for a target object [76,78]. Another study [70] predicted fixation scanpaths during search using COCO-Search18 [13], a dataset of search fixations. Using a dataset [10] of fixation scanpaths from a Visual Question Answering (VQA) task, Chen *et al*. [11] proposed a model that predicted both VQA and search behavior, and both Chen *et al*. [14], Yang *et al*. [72] proposed models to predict both target-present and target-absent search. Mondal *et al*. [50] proposed a multimodal transformer model called Gazeformer, which achieves state-of-the-art search prediction performance while generalizing well to unknown targets. Gazeformer [50] and Chen *et al*. [11] can be adapted for our multimodal task, and serve as baselines in Sec. 5.

Despite this increasing interest in gaze prediction as a computer vision problem, no existing model effectively addresses the incremental object referral task. Large vision-language foundation models [2,26,55,67,73,75] yield unprecedented

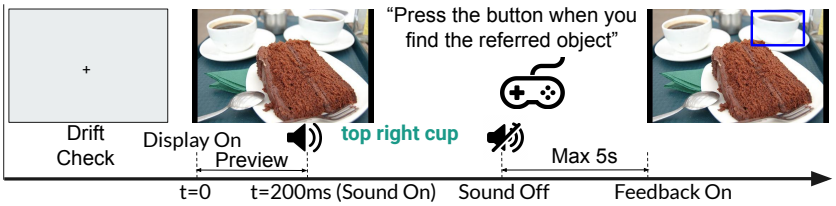

Fig. 2: Behavioral data collection in our incremental object referral paradigm.

performance in visual, lingual and cross-modal tasks and effectively generalize to new concepts and tasks. Moreover, several tasks simply require multimodal modeling, such as VQA [4,48], image captioning [19,34], and object referral [45,69,74]. Relatedly, *object referral* (also known as visual/object grounding of referring expressions) localizes or *grounds* a single unambiguous object in an image that is referred to in a natural language *referring expression*. Thus, the input is an image-text pair and the output is the referred object's bounding box parameters. Recent object referral models have adopted two-stage [23, 24, 43] and one-stage [12,17,41] architectures, and to this end, several high-quality object referral benchmarks have been curated, such as ReferItGame [30], RefCOCO [74], RefCOCO+ [74], and RefCOCOg [45]. Researchers have also studied human attention as people view an image and concurrently describe it [62,63]. One study [54] collected a dataset of spoken image descriptions where each word was visually grounded by a mouse trace. He *et al*. [21] collected a dataset containing fixations (recorded by an eye-tracker) synchronized with concurrently spoken image descriptions. However, these studies specifically focused on *spoken description of the entire image* and not object referral. Vasudevan *et al*. [65] explored object referral for previously spoken referring expressions, and did not predict human attention. Another study [64] on spoken object referral in videos used human gaze and spoken referring expression as inputs. Zhang *et al*. [79] collected a dataset of static gaze estimation heatmaps for non-incremental referral. To our knowledge, we are the first to computationally model human gaze and explore its interactions with vision and language in a realistic incremental referral task.

3 RefCOCO-Gaze Dataset

RefCOCO-Gaze is the largest dataset for studying human gaze behavior during an incremental object referral task. It consists of 19,738 scanpaths that were recorded while 220 participants with normal or corrected-to-normal vision viewed 2,094 COCO [42] images and listened to the associated referring expressions from the RefCOCO dataset [74]. RefCOCO was collected using the ReferItGame [30] where players must construct efficient referring expressions for another player to locate the correct object. RefCOCO mirrors real-life speech which is known to contain elliptical and unstructured expressions [60, 61]. The gaze data, recorded by an EyeLink 1000 eyetracker, includes information about the location and duration of each fixation, the bounding box of the search target, audio recordings of

the referring expressions, the timing of the target word, and the synchronization between the spoken words and the sequence of fixations (tells us which word triggered which fixations). RefCOCO-Gaze covers a diverse range of linguistic and visual complexity, making it an ideal dataset for researchers studying human integration of vision and language, and HCI researchers alike.

Fig. 2 depicts the incremental object referral paradigm used for human gaze collection. We have selected 2,094 image-expression pairs from the larger RefCOCO dataset based on the target object size, image ratio, and sentence complexity. Each participant performed ∼100 trials, yielding 10-16 scanpaths per image-expression pair. Participants were instructed to move their gaze as quickly as possible to the target object that is being referred to in a language expression played through a speaker. Each trial began with drift correction (for accurate eye tracking) and presentation of an image. The image was displayed for 200ms before the audio onset (too short a time for an eye movement). The image remained visible until the participant pressed a button to indicate that they found the target or until five seconds elapsed following completion of the auditorily presented referring expression. At the end of each trial, the correct bounding box location of the target object was provided as feedback, followed by a survey asking whether the participant indeed found and recognized the target. A target was present in each image. We used a forced aligner [47], a tool for aligning speech with text, to synchronize gaze movements with individual words of a referring expression. This study had IRB approval.

We divided RefCOCO-Gaze into disjoint training and evaluation sets that preserve the approximate proportion of training to evaluation data in RefCOCO. The training set consisted of 1799 image-expression pairs (selected *only* from the original RefCOCO train split), corresponding to 16,982 scanpaths. Scanpaths from image-expression pairs from validation and test splits of RefCOCO were randomly shuffled and split (1:2 ratio) to create disjoint validation (92 image-expression pairs, 869 scanpaths) and test (203 image-expression pairs, 1887 scanpaths) sets – both having a balanced distribution of target categories. Dataset details (*e.g.*, stimuli selection, gaze recording, pre-processing, comparison with related datasets, etc.) and extensive dataset analyses are in the supplement.

4 Gaze Prediction for Incremental Object Referral

Our goal is to predict fixations as a person progressively receives information about the referred object through *each word* of the referring expression that they are hearing. We design a novel multimodal transformer architecture called ***A**ttention in **R**eferral **T**ransformer* or **ART** for this task. ART solves multiple problems that arise when adapting previous gaze prediction models [11, 50] for the incremental object referral task, owing to its several novel features that these models [11, 50] lack: **(1)** ART integrates an object referral network into our gaze prediction framework and trains it on partial expressions, whereas previous baselines either extract task guidance from a frozen object referral model [28] trained on complete expressions [11], or lack object grounding capabilities [50].

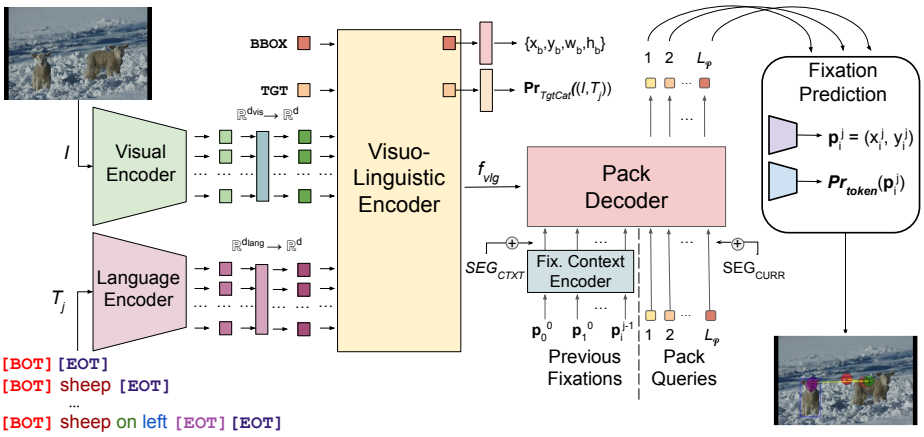

Fig. 3: Attention in Referral Transformer (ART) Architecture. On each pass after comprehending a new word, the model takes an image I and tokens T_j of prefix R_j of the referring expression as input and generates a possibly empty sub-sequence of fixations based on previous fixation history encoded by a fixation context encoder.

(2) ART integrates a novel fixation prediction framework (absent in previous methods) that accommodates the autoregressive prediction of any number of fixations triggered by a word – zero or one or several, based on previous fixations encoded by a novel fixation context encoder (in contrast to non-autoregressive Gazeformer [50] which predicts fixations with the knowledge of the first fixation alone). **(3)** ART is an object referral network with a scanpath prediction decoder module, which allows us to *pre-train* the object referral network on a large-scale object referral dataset [74] using object grounding objectives, thus generalizing better despite training on a much smaller gaze dataset. ART also jointly trains on the object grounding objectives along with the primary fixation prediction objective. On the other hand, previous baselines either rely on frozen object referral models [11] or lack an object grounding subnetwork altogether [50] and are therefore limited to only training on the fixation prediction objective alone.

4.1 Architecture

The overall architecture of ART is shown in Fig. 3. Since the understanding of the referred object changes with each incoming word, we design ART to output a possibly empty sequence of fixations (which we dub a "pack" of fixations) triggered by a new word w_j from the referring expression. We follow single-stream vision-language model architectures [17, 38–40] to design a multimodal transformer encoder module that encodes visuo-linguistic context for object referral. A transformer decoder module generates embeddings for incremental scanpath prediction conditioned by visuo-linguistic context from the encoder and fixation history encoded by a fixation context encoder module.

Visual and Language Encoders. We use separate encoders for the vision and language modalities. As in DETR [9], the **visual encoder** consists of a ResNet-50 [20] backbone followed by a standard transformer encoder module [66]. Given an image $I \in \mathbb{R}^{3\times H\times W}$, the visual encoder generates patch embeddings $g_{vis} \in \mathbb{R}^{d_{vis}\times hw}$. A **language encoder** (RoBERTa [44]) encodes a *prefix* $R_j = \{w_1, \ldots, w_j\}$ upon utterance of the j^{th} word w_j in a referring expression. R_j is tokenized [57] to obtain $T_j = \{\texttt{[BOT]}, t_1, ..., \texttt{[EOT]}\}$, which is then processed by the language encoder to yield language embedding sequence $g_{lang} \in \mathbb{R}^{d_{lang}\times l_{lang}}$. Here, l_{lang} is the maximum number of tokens that can be processed by the language model; `[BOT]` and `[EOT]` are the beginning-of-text and end-of-text tokens, respectively. To predict fixations triggered *before* the first word has been spoken, the input tokens to the language encoder is the sequence {`[BOT]`, `[EOT]`}, and to predict fixations *after* the expression has ended, we append an additional `[EOT]` token to the tokenized input. Existing baselines [11,50] use ResNet-50 and RoBERTa, hence we choose them for fair comparison.

Visuo-linguistic Transformer Encoder. In contrast to Gazeformer [50], which utilized pooled text encodings from a frozen language encoder for task description and risked losing word-level details, our method integrates both image and linguistic tokens into a unified sequence of multimodal tokens, thus allowing fine-grained word-level interactions between image and linguistic tokens. We process this sequence through a visuo-linguistic encoder, which is designed as a standard transformer encoder [66]. Given the set of patch embedding vectors g_{vis} and the sequence of language embedding vectors g_{lang} with different dimensionalities, as described above, we first project them to an embedding space of the same dimension d using modality-specific projections to obtain $f_{vis} \in \mathbb{R}^{d\times hw}$ and $f_{lang} \in \mathbb{R}^{d\times l_{lang}}$. We also introduce two learnable d-dimensional embeddings `BBOX`, `TGT` which correspond to the object localization and target category prediction tasks, respectively. The input to the visuo-linguistic encoder is the concatenation of `BBOX`, `TGT`, f_{vis}, and f_{lang}. The corresponding output is tensor $f_{vlg} \in \mathbb{R}^{d\times(hw+l_{lang}+2)}$. We project $f_{vlg}[0]$ (corresponding to `BBOX`) to $\{x_b, y_b, w_b, h_b\}$ where x_b, y_b are the coordinates of the upper-left corner, and w_b and h_b are the width and height of the bounding box. We also use a linear layer along with softmax to project $f_{vlg}[1]$ (corresponding to `TGT`) to the probability distribution $\mathbf{Pr}_{TgtCat}$ over all possible target categories.

Per-Word Fixation Prediction Framework. We predict fixations (or absence thereof) triggered by a new word of a referring expression containing L words. Let the scanpath $\mathcal{S}$ for incremental object referral be a sequence of *packs* of fixations. A pack $\mathcal{P}_j = \{(x_i^j, y_i^j) | i = 0, 1, \ldots\}$ is an ordered sequence of 2D fixations, triggered by the change in knowledge about the referred object due to a new word w_j, for $j \in \{1, .., L\}$. While a pack is usually spurred by a word, a pack of fixations $\mathcal{P}_0$ can be triggered *before* the first word is spoken, similar to free-viewing behavior. A pack of fixations $\mathcal{P}_{L+1}$ can also occur *after* the referring expression has ended. A word may not inspire any fixations at all, yielding a *null* pack $\mathcal{P}_\phi = \phi$. We also define a *terminal* pack $\mathcal{P}_{TERM}$ which, like the null pack, does not contain any valid fixations, and denotes the end of scanpath (EOS).

Fixation context encoder. We parameterize a fixation $\mathbf{p}_i^k$ using four parameters: x-location x_i^k, y-location y_i^k, the pack number k (i.e., the index of the pack the fixation belongs to), and the within-pack index i (which we call *order*). We use this parameterization to capture the *fixation context*, which refers to the information of previous fixations in the ongoing scanpath. We use a fixed 2D sinusoidal positional embedding [9] to encode the spatial x, y location to $XY_i^k \in \mathbb{R}^{2d}$ and two fixed 1D sinusoidal positional embeddings [66] to encode the pack number j and the order i to $n_i^k, o_i^k \in \mathbb{R}^d$ respectively. XY_i^k, n_i^k, and o_i^k are concatenated and projected to fixation encoding $\mathbf{c}_i^k \in \mathbb{R}^d$. Hence, for a new word w_j, we can construct an ordered sequence of $\mathbf{c}_i^k$ $(k < j, i = 0, 1, ...)$ and zero-pad to maximum length $L_\mathcal{C}$ to obtain the fixation context tensor $\mathcal{C}_j \in \mathbb{R}^{d \times L_\mathcal{C}}$.

Pack decoder. To obtain the current pack of fixations, we use a transformer decoder module [66]. Let the input $\mathcal{Q} = \{q_k | k = 1, \ldots, L_\mathcal{P}\}$ to the decoder be a sequence of pack queries q_k, where $L_\mathcal{P}$ is the maximum number of fixations in a pack and q_k's are learnable vectors (similar to fixation queries [50]). To help the model differentiate the nature of context and pack embeddings, we further add two separate segment embeddings [18], namely SEG_{ctxt} and SEG_{curr} to previous fixation context tensor $\mathcal{C}_j$ and $\mathcal{Q}$, respectively. Next, we use the concatenation of $\mathcal{C}_j$ and $\mathcal{Q}$ as input to the decoder. The decoder also receives dynamic visuo-linguistic context through cross-attention with f_{vlg}. The output from the decoder is tensor $f_{decoder} \in \mathbb{R}^{d \times (L_\mathcal{C} + L_\mathcal{P})}$. The last $L_\mathcal{P}$ d-dimensional slices of $f_{decoder}$ corresponding to the $L_\mathcal{P}$ pack queries are denoted as $f_{pack} \in \mathbb{R}^{d \times L_\mathcal{P}}$.

Fixation Prediction Module. Fixation prediction for incremental object referral is challenging since a pack can have between zero and multiple fixations, and a scanpath can be terminated before the end of a referring expression. We account for these scenarios by making all packs be of length $L_\mathcal{P}$ with the following parameterization. First, any valid fixation in a pack is represented by a fixation token `FIX`. Second, null packs and packs having less than $L_\mathcal{P}$ valid fixations are padded with padding tokens `PAD` to maximum length $L_\mathcal{P}$. Third, we complete a terminal pack $\mathcal{P}_{TERM}$ with $L_\mathcal{P}$ termination tokens `EOS`. For each of the $L_\mathcal{P}$ slices of f_{pack}, we use a token prediction MLP and a softmax layer to predict if that slice corresponds to one of `FIX`, `PAD`, and `EOS` tokens. We use regression heads and Gaussian distributions [50] to model the fixation locations. We also augment ART with fixation duration modeling and detail it in the supplement.

4.2 Pre-training, Training, and Inference

Pre-training. Since object grounding is at the core of our task, we *pre-train* the visual, language and visuo-linguistic encoder modules on the two objectives that we hypothesize underlie the object grounding process: object localization and target category prediction, using RefCOCO [74] training data. **Object localization** is the estimation of the referred object bounding box. We apply an L_1 regression loss $\mathcal{L}_{reg}$ and a generalized IoU (GIoU) loss [56] $\mathcal{L}_{giou}$, between predicted and ground truth bounding box parameters. The **target category prediction**

task discerns the object type from the expression (e.g., predict "car" in the expression "left sedan next to the motorcycle"). We pre-train on this task using a cross-entropy loss $\mathcal{L}_{target}$. Total pre-training loss is $\mathcal{L}_{pretrain} = \mathcal{L}_{reg} + \mathcal{L}_{giou} + \mathcal{L}_{target}$.

Training & Inference. We train ART using the teacher-forcing algorithm [68], i.e., we provide the ground truth fixations to construct the fixation context and treat each pack in the training scanpaths as independent minibatch items. To train ART on the gaze prediction task, we apply $\mathcal{L}_1$ regression loss (following [50]) on the predicted x and y locations. Let the predicted pack of fixations $\mathcal{P}_k = \{(x_i^k, y_i^k)\}_{i=1}^{L_\mathcal{P}}$, and ground-truth pack of fixations $\hat{\mathcal{P}}_k = \{(\hat{x}_i^k, \hat{y}_i^k)\}_{i=1}^{l^k}$ where l^k is the length of the ground truth pack. Moreover, let $\hat{v}_{i,t}^k$ be a binary scalar representing ground truth of the i^{th} token in $\mathcal{P}_k$ belonging to the token class $t \in T$ where $T = \{\texttt{FIX}, \texttt{PAD}, \texttt{EOS}\}$. Also let $v_{i,t}^k$ be the probability of that token belonging to token class t as estimated by our model. The multitask loss for a minibatch of size M is

$$\mathcal{L}_{gaze} = \frac{1}{M} \sum_{k=1}^{M} \left(\mathcal{L}_{xy}^k + \mathcal{L}_{token}^k \right). \tag{1}$$

Here $\mathcal{L}_{xy}^k = \frac{1}{l^k} \sum_{i=1}^{l^k} \left(|x_i^k - \hat{x}_i^k| + |y_i^k - \hat{y}_i^k| \right)$, $\mathcal{L}_{token}^k = -\sum_{i=1}^{L_\mathcal{P}} \sum_{t \in T} \hat{v}_{i,t}^k \log(v_{i,t}^k)$. In addition to the gaze-prediction loss $\mathcal{L}_{gaze}$, we also train on the object localization and target category prediction tasks, but only *after* either of the following two events has occurred: (1) the last word of the referring expression has been uttered, (2) the ground truth scanpath has been terminated. Note that both events ensure sufficient information in the referring expression comprehended thus far for a human to localize the object. This multi-task grounding loss $\mathcal{L}_{ground}$ is $\mathcal{L}_{bbox} + \mathcal{L}_{target}$, where $\mathcal{L}_{bbox} = \mathcal{L}_{reg} + \mathcal{L}_{giou}$. Hence, the total multi-task loss $\mathcal{L}$ that we use to train our ART model is $\mathcal{L} = \mathcal{L}_{gaze} + \mathcal{L}_{ground}$ when the scanpath has terminated or the referral audio has ended, and $\mathcal{L} = \mathcal{L}_{gaze}$ otherwise. During inference, ART *autoregressively* generates packs of fixations conditioned on the previous fixations generated by the model and the scanpath is terminated upon encountering the first termination token `EOS` in a predicted pack. The fixations within a pack are efficiently generated in parallel.

5 Experiments

Here, we experimentally evaluate scanpath prediction capability for incremental object referral. For the conventional scanpath prediction task, accurately predicting the entire sequence of fixations is the main objective. However, for our task, it is perhaps equally important, if not more, for the predicted scanpath to be correct at the word-level granularity, i.e., packs (including null packs) must be predicted accurately. Following previous work [11, 50, 70, 72], we sample 10 scanpaths per image-expression pair for all models. More details of ART, such as its design and implementation details, are in the supplement.

5.1 Performance Metrics

We use a broad set of metrics to evaluate dynamic word-based scanpath prediction for incremental object referral. *Sequence Score* metric [70] converts predicted and ground truth scanpaths into strings of fixation cluster IDs and compares them using a string matching algorithm [51]. *Fixation Edit Distance* [50] measures scanpath dissimilarity using the Levenshtein algorithm [37] after converting scanpaths to strings like Sequence Score does. We measure Sequence Score and Fixation Edit Distance in two granularities: (1) over the entire scanpath (SS and FED); and (2) over a pack (SS_{pack} and FED_{pack}), where SS_{pack} and FED_{pack} are the averages of sequence scores and fixation edit distances, respectively, between the ground-truth and predicted packs. We also introduce CC_{pack} and NSS_{pack}, the word-based versions of the *Correlation Coefficient (CC) [27]* and *Normalized Scanpath Saliency (NSS) [53]* metrics. CC is the correlation between the normalized model saliency map and a Gaussian-convolved human fixation map. NSS averages the values of a model's fixation map at the locations fixated by humans [8], and is a discrete version of CC. CC_{pack} and NSS_{pack} are the averages of CC scores and NSS scores, respectively, over all possible packs. Higher SS, SS_{pack}, NSS_{pack}, and CC_{pack} values signify more similarity between model-generated and human scanpaths, whereas lower FED and FED_{pack} scores indicate higher similarity. More details are in the supplement.

5.2 Baselines

We compare ART with: (1) **Random Scanpath**: We uniformly sample a pack length value l_p and then uniformly sample l_p fixation locations from the image. (2) **OFA**: We use the state-of-the-art vision-language model OFA [67], trained on several multimodal benchmarks. We uniformly sample pack length l_p and then sample l_p fixation locations within the OFA-predicted *bounding box* for each referring expression prefix. (3) **Chen *et al.*** [11]: This model learns goal-directed human gaze through a dynamically updated memory which is initialized by task guidance maps. To extend this model to our task, we create task guidance maps using bounding boxes from the SOTA referral model MDETR [28], trained only on RefCOCO. (4) **Gazeformer-ref**: This baseline, based on Gazeformer [50], takes expression prefixes as target information and generates packs of fixations. (5) **Gazeformer-cat**: Since target category information might get lost in the pooled linguistic embedding used by Gazeformer-ref, we evaluate another variant of Gazeformer [50] called *Gazeformer-cat* which takes the *target category name* estimated for an expression prefix as input and treats the problem as categorical visual search. The target category estimation of a prefix is done by a pre-trained RoBERTa-based classifier. Find more details of the baselines in the supplement.

5.3 Results

We train ART and the baselines on the RefCOCO-Gaze training set and evaluate them on the test set. Results are in Table 1. ART outperforms baselines on all

Table 1: Performance of ART and baselines on RefCOCO-Gaze test set.

	$SS \uparrow$	$SS_{pack} \uparrow$	$FED \downarrow$	$FED_{pack} \downarrow$	$CC_{pack} \uparrow$	$NSS_{pack} \uparrow$
Human	0.400	0.317	6.573	1.278	0.283	3.112
Random	0.189	0.133	17.735	3.005	0.094	1.689
OFA [67]	0.216	0.170	17.084	2.901	0.174	2.175
Chen *et al.* [11]	0.299	0.188	8.309	1.507	0.159	1.557
Gazeformer-ref [50]	0.269	0.194	6.788	1.286	0.208	3.006
Gazeformer-cat [50]	0.269	0.189	6.841	1.327	0.204	2.932
ART (Proposed)	**0.359**	**0.292**	**6.371**	**1.143**	**0.280**	**3.478**

metrics by significant margins. We hypothesize that ART performs best because it *includes* an object referral model and jointly trains on grounding and gaze prediction objectives. In contrast, Chen *et al.* [11] use a frozen MDETR model trained only on complete RefCOCO expressions, and the Gazeformer variants (Gazeformer-ref and Gazeformer-cat) lack grounding subnetworks to train on object grounding. Hence, both baselines are unable to learn the auxiliary grounding tasks on partial expressions. Interestingly, despite using no spatial and attribute information, Gazeformer-cat is almost as predictive as Gazeformer-ref, underscoring the importance of target category estimation. Note that these analyses focused on spatial attention, but in the supplement, we show that ART also outperforms baselines in terms of fixation duration prediction as well. We also show in the supplement that ART *generalizes well to categorical search* when trained and evaluated using COCO-Search18 [13] dataset.

Fig. 4 shows qualitative results comparing the sequences of fixations from ART and the baselines to the behavioral data. In these three examples, ART finds the referred object and generates efficient, human-like scanpaths. ART also exhibits several strategic fixation patterns that we observe in the human data. In the top row, ART *waits* near the center until after getting the word "right", which conveys information about the referred sheep. A *scanning* gaze pattern appears in the second row, where both the person and ART scan multiple bags, thereby enabling the correct one to be located when the disambiguating information arrives at the end of the expression. The third row exemplifies *verification*. ART successfully finds the correct target on fixation #3 after input of the word "girl", but then makes another fixation (#5) to the girl in the center after getting the word "pink", presumably to verify which of the girls is pinker before returning to the one on the left on the next fixation (#6).

5.4 Ablation Studies

We performed a number of ablations (in Table 2) on ART to probe the effects of pre-training and inclusion of grounding losses on its performance, which we evaluate using the RefCOCO-Gaze test set. As evidenced by comparing Ablations 4 and 5, pre-training significantly improves performance. We also observe that

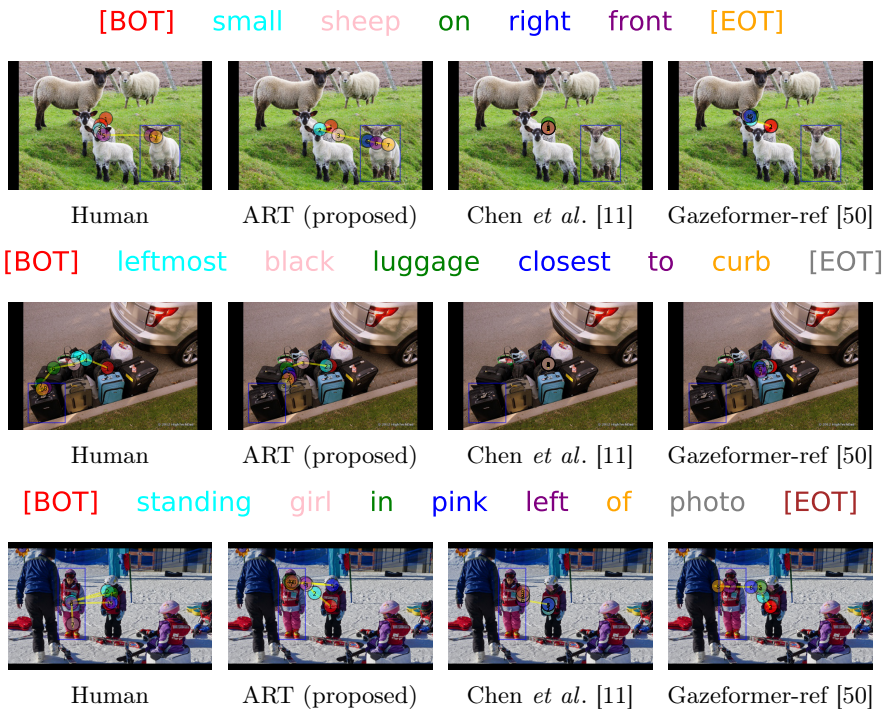

Fig. 4: Qualitative results. Scanpaths from humans and three scanpath prediction models on three trials exhibiting strategic fixation behavior. Fixations (denoted by circles numbered with fixation order) are color-coded to corresponding words in the referring expression (above each row). Fixations color-coded to `[BOT]` occurred before the expression started, and those color-coded to `[EOT]` occurred after the expression ended. Blue bounding boxes indicating the referred objects are not visible during trials. Our model generates the most human-like scanpaths for incremental object referral.

without pre-training, the grounding losses $\mathcal{L}_{bbox}$ and $\mathcal{L}_{target}$ slightly degrade performance (compare Ablations 1 and 4). When trained from scratch on the small gaze dataset, these auxiliary tasks introduce noise to the optimization of the gaze prediction objective using $\mathcal{L}_{gaze}$. Including one of $\mathcal{L}_{bbox}$ (Ablation 2) and $\mathcal{L}_{target}$ (Ablation 3) losses in pre-training and training does not improve performance significantly (although $\mathcal{L}_{bbox}$ seems to be more beneficial than $\mathcal{L}_{target}$), whereas including *both* $\mathcal{L}_{bbox}$ and $\mathcal{L}_{target}$ (main model and Ablation 5) yields the best performance. This demonstrates that both object localization and target category estimation tasks are integral to the object referral process. We note that even when ART is not pre-trained on RefCOCO (Ablations 1, 4), it still outperforms baselines like the two Gazeformer variants that are also not pre-trained on RefCOCO. See the supplement for more metrics, ablations and analyses.

Table 2: Ablation studies on ART model. If either $\mathcal{L}_{bbox}$ or $\mathcal{L}_{target}$ is included, and the model undergoes pre-training, the loss is applied in *both* pre-training and gaze training phases. Additional metrics and more ablation studies are in the supplement.

Ablation #	Pre-training	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	$SS \uparrow$	$SS_{pack} \uparrow$	$CC_{pack} \uparrow$
1	×	×	×	0.309	0.257	0.222
2	✓	✓	×	0.321	0.279	0.239
3	✓	×	✓	0.292	0.260	0.216
4	×	✓	✓	0.304	0.257	0.215
5	✓	✓	✓	**0.359**	**0.292**	**0.280**

6 Conclusions and Discussion

How do humans integrate vision and language information to guide their attention to target goals? To study this question we introduced the *incremental object referral* task, a naturalistic version of an object referral task in which people must incrementally integrate the visual information that they are actively collecting with each location that they fixate in the image, with the language information that they are hearing about the target object. Our task therefore provides an experimental context for studying how humans use the spoken language of another to dynamically control the visual information that they sample from the world. We also introduced a model that we call *ART* that similarly generates sequences of gaze fixations that occur as this vision-language integration is happening. ART has a multimodal transformer architecture that it uses to learn how to incrementally generate *packs* of fixations for each word in the referring expression. To provide the human behavior needed to train ART, we collected a high-quality and large-scale dataset called *RefCOCO-Gaze*. We trained and evaluated ART and several competitive baselines on RefCOCO-Gaze and found that ART outperformed other baselines by significant margins on multiple metrics. We also performed extensive ablation analyses to show how pre-training ART on RefCOCO, and the addition of auxiliary grounding losses, significantly contributed to its superior performance. Qualitative analysis revealed that ART showed human-like effects of visual and linguistic target ambiguity on its attention behavior through higher-level strategic forms of integrating vision and language information, expressed as distinct *waiting*, *verification*, and *scanning* strategies. We believe that ART will be instrumental for predicting gaze in time-sensitive, voice-assisted HCI applications (especially AR/VR) where predicting future eye movements will enable seamless human-computer interactions.

A current limitation of ART is that it treats the referring expression as text and not audio, thereby ignoring the phonological factors influencing vision-language disambiguation and attention control. Future work will explore representing these phonological factors as well.

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

References

1. Adhanom, I.B., Griffin, N.N., MacNeilage, P., Folmer, E.: The effect of a foveated field-of-view restrictor on vr sickness. In: 2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE (2020)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems (2022)
3. Altmann, G.T.: Language can mediate eye movement control within 100milliseconds, regardless of whether there is anything to move the eyes to. Acta Psychologica **137**(2), 190–200 (Jun 2011)
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual Question Answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2015)
5. Bapna, T., Valles, J., Leng, S., Pacilli, M., Nataraja, R.M.: Eye-tracking in surgery: a systematic review. ANZ Journal of Surgery **93**(11), 2600–2608 (2023)
6. Bennett, C.R., Bex, P.J., Merabet, L.B.: Assessing visual search performance using a novel dynamic naturalistic scene. Journal of Vision **21**(1), 5–5 (2021)
7. Berg, D.J., Boehnke, S.E., Marino, R.A., Munoz, D.P., Itti, L.: Free viewing of dynamic stimuli by humans and monkeys. Journal of vision **9**(5), 19–19 (2009)
8. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(3), 740–757 (2019)
9. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision (2020)
10. Chen, S., Jiang, M., Yang, J., Zhao, Q.: Air: Attention with reasoning capability. In: European Conference on Computer Vision (2020)
11. Chen, X., Jiang, M., Zhao, Q.: Predicting human scanpaths in visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
12. Chen, X., Ma, L., Chen, J., Jie, Z., Liu, W., Luo, J.: Real-time referring expression comprehension by single-stage grounding network. arXiv preprint arXiv:1812.03426 (2018)
13. Chen, Y., Yang, Z., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G.: Coco-search18 fixation dataset for predicting goal-directed attention control. Scientific reports **11**(1), 8776 (2021)
14. Chen, Y., Yang, Z., Chakraborty, S., Mondal, S., Ahn, S., Samaras, D., Hoai, M., Zelinsky, G.: Characterizing target-absent human attention. In: Proceedings of CVPR International Workshop on Gaze Estimation and Prediction in the Wild (2022)
15. Chung, J., Lee, H., Moon, H., Lee, E.: The static and dynamic analyses of drivers' gaze movement using vr driving simulator. Applied Sciences **12**(5), 2362 (2022)
16. Cooper, R.M.: The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. Cognitive Psychology **6**(1), 84–107 (1974)
17. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)

18. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (2019)
19. Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J.C., et al.: From captions to visual concepts and back. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2015)
20. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)
21. He, S., Tavakoli, H.R., Borji, A., Pugeault, N.: Human attention in image captioning: Dataset and analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
22. Henderson, J.M., Brockmole, J.R., Castelhano, M.S., Mack, M.: Visual saliency does not account for eye movements during visual search in real-world scenes. In: Eye movements, pp. 537–III. Elsevier (2007)
23. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(2), 684–696 (2019)
24. Hu, R., Rohrbach, M., Andreas, J., Darrell, T., Saenko, K.: Modeling relationships in referential expressions with compositional modular networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
25. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **20**(11), 1254–1259 (1998)
26. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning (2021)
27. Jost, T., Ouerhani, N., Von Wartburg, R., Müri, R., Hügli, H.: Assessing the contribution of color in visual attention. Computer Vision and Image Understanding **100**(1-2), 107–123 (2005)
28. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
29. Kamide, Y., Altmann, G.T., Haywood, S.L.: The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. Journal of Memory and language **49**(1), 133–156 (2003)
30. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
31. Khokhar, A., Yoshimura, A., Borst, C.: Eye-gaze-triggered visual cues to restore attention in educational vr. In: 2019 IEEE conference on virtual reality and 3D user interfaces (VR), poster (2019)
32. Knoeferle, P., Guerra, E.: Visually Situated Language Comprehension: Visually Situated Language Comprehension. Language and Linguistics Compass **10**(2), 66–82 (Feb 2016)
33. Koehler, K., Guo, F., Zhang, S., Eckstein, M.P.: What do saliency models predict? Journal of vision **14**(3), 14–14 (2014)

34. Kuo, C.W., Kira, Z.: Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022)
35. Lang, Y., Wei, L., Xu, F., Zhao, Y., Yu, L.F.: Synthesizing personalized training programs for improving driving habits via virtual reality. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). IEEE (2018)
36. Lavoie, E., Hebert, J.S., Chapman, C.S.: Comparing eye–hand coordination between controller-mediated virtual reality, and a real-world object interaction task. Journal of Vision **24**(2), 9–9 (2024)
37. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet physics. Doklady **10**, 707–710 (1965)
38. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
39. Li, P., Tian, B., Shi, Y., Chen, X., Zhao, H., Zhou, G., Zhang, Y.Q.: Toist: Task oriented instance segmentation transformer with noun-pronoun distillation. Advances in Neural Information Processing Systems (2022)
40. Li, Y., Chen, X., Zhao, H., Gong, J., Zhou, G., Rossano, F., Zhu, Y.: Understanding embodied reference with touch-line transformer. In: International Conference on Learning Representations (2023)
41. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
42. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision. pp. 740–755. Springer (2014)
43. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)
44. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
45. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2016)
46. Masciocchi, C.M., Mihalas, S., Parkhurst, D., Niebur, E.: Everyone knows what is interesting: Salient locations which should be fixated. Journal of vision **9**(11), 25–25 (2009)
47. McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., Sonderegger, M.: Montreal forced aligner: Trainable text-speech alignment using kaldi. In: Interspeech (2017)
48. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2023)
49. Min, K., Corso, J.J.: Integrating human gaze into attention for egocentric activity recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2021)
50. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In: Pro-

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
51. Needleman, S.B., Wunsch, C.D.: 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)
52. Pai, Y.S., Tag, B., Outram, B., Vontin, N., Sugiura, K., Kunze, K.: Gazesim: Simulating foveated rendering using depth in eye gaze for vr. In: ACM SIGGRAPH 2016 Posters (2016)
53. Peters, R.J., Iyer, A., Koch, C., Itti, L.: Components of bottom-up gaze allocation in natural scenes. Journal of Vision **5**(8), 692–692 (2005)
54. Pont-Tuset, J., Uijlings, J., Changpinyo, S., Soricut, R., Ferrari, V.: Connecting vision and language with localized narratives. In: European Conference on Computer Vision (2020)
55. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning (2021)
56. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
57. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016)
58. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C.: Integration of Visual and Linguistic Information in Spoken Language Comprehension. Science **268**(5217), 1632–1634 (Jun 1995)
59. Tanenhaus, M.K., Spivey-Knowlton, M.J., Eberhard, K.M., Sedivy, J.C.: Using eye movements to study spoken language comprehension: Evidence for visually mediated incremental interpretation. (1996)
60. Thanh, N.C.: The differences between spoken and written grammar in english, in comparison with vietnamese (las diferencias entre la gramática oral y escrita del idioma inglés en comparación con el idioma vietnamita). Gist Education and Learning Research Journal **11**, 138–153 (2015)
61. Townend, J., Walker, J.: Structure of Language: Spoken and Written English. Whurr Publishers (2006)
62. Vaidyanathan, P., Prud'hommeaux, E., Alm, C.O., Pelz, J.B.: Computational framework for fusing eye movements and spoken narratives for image annotation. Journal of Vision **20**(7), 13–13 (2020)
63. Vaidyanathan, P., Prud'hommeaux, E., Pelz, J.B., Alm, C.O.: Snag: Spoken narratives and gaze dataset. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (2018)
64. Vasudevan, A.B., Dai, D., Van Gool, L.: Object referring in videos with language and human gaze. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
65. Vasudevan, A.B., Dai, D., Van Gool, L.: Object referring in visual scene with spoken language. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2018)
66. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (2017)

67. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: International Conference on Machine Learning (2022)
68. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation **1**(2), 270–280 (1989)
69. Yan, B., Jiang, Y., Wu, J., Wang, D., Luo, P., Yuan, Z., Lu, H.: Universal instance perception as object discovery and retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
70. Yang, Z., Huang, L., Chen, Y., Wei, Z., Ahn, S., Zelinsky, G., Samaras, D., Hoai, M.: Predicting goal-directed human attention using inverse reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
71. Yang, Z., Mondal, S., Ahn, S., Xue, R., Zelinsky, G., Hoai, M., Samaras, D.: Unifying top-down and bottom-up scanpath prediction using transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
72. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., Samaras, D.: Target-absent human attention. In: European Conference on Computer Vision (2022)
73. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. Transactions on Machine Learning Research (2022)
74. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision (2016)
75. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
76. Zelinsky, G., Yang, Z., Huang, L., Chen, Y., Ahn, S., Wei, Z., Adeli, H., Samaras, D., Hoai, M.: Benchmarking gaze prediction for categorical visual search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
77. Zelinsky, G.J., Chen, Y., Ahn, S., Adeli, H.: Changing perspectives on goal-directed attention control: The past, present, and future of modeling fixations during visual search. In: Psychology of Learning and Motivation, vol. 73, pp. 231–286. Elsevier (2020)
78. Zelinsky, G.J., Chen, Y., Ahn, S., Adeli, H., Yang, Z., Huang, L., Samaras, D., Hoai, M.: Predicting goal-directed attention control using inverse-reinforcement learning. Neurons, behavior, data analysis and theory (2020)
79. Zhang, D., Tian, Y., Chen, K., Qian, K.: Gaze-directed visual grounding under object referring uncertainty. In: 2022 41st Chinese Control Conference (CCC). IEEE (2022)