Supplementary Material for 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

In this document, we provide additional experiments, visualizations and details of our work on gaze prediction during *incremental object referral* task using the *Attention in Referral Transformer (ART)* model and *RefCOCO-Gaze* dataset. The specific sections of this document are listed below.

- We provide details about how we selected the stimuli, i.e. the images and referring expressions, from RefCOCO dataset to create RefCOCO-Gaze dataset. (Section 1).
- We discuss the gaze recording method we used to collect human fixations for RefCOCO-Gaze dataset along with analysis of the collected gaze data (Section 2).
- We provide a comparative analysis of our proposed RefCOCO-Gaze dataset and other related gaze datasets discussed in the main text (Section 3).
- We provide details of various components of the ART model along with the pre-training and training procedures (Section 4).
- We provide implementation details of several scanpath metrics used for evaluation in the main text (Section 5).
- We augment ART with fixation duration prediction capability and report the experimental results for ART model and other baselines on RefCOCO-Gaze with respect to both fixation location and fixation duration prediction (Section 6).
- We show that ART generalizes to categorical search task when trained and evaluated using COCO-Search18 [4] dataset (Section 7).
- We provide implementation details of the baselines - Random Scanpath, OFA [27], Chen *et al*. [3], Gazeformer-ref [16], and Gazeformer-cat [16] (Section 8).
- We augment the experimental results for the ablation studies on ART model, which are discussed in the main text, with additional metrics (Section 9).
- We present additional ablation studies investigating the effects that the components of ART model have on performance and also include additional analysis on the ablations (Section 10).
- We explore the effects of next word token prediction task on the performance of ART model (Section 11).

– We present additional qualitative examples of scanpaths generated by human participants, our ART model and other competitive baseline models. (Section 12).

1 Image and Referring Expression Selection Details

We utilized a subset of the RefCOCO dataset [32] (the original UNC split) to create our dataset. RefCOCO dataset consists of referring expressions collected for 50,000 target objects present in 19,994 COCO [12] training images. RefCOCO was carefully curated such that each image contained at least two objects of the same object category as the target object. To ensure the reliability of our gaze data, we selected the longest referring expression amongst the multiple referring expressions collected for each target object. To eliminate stimuli that might produce inaccurate gaze patterns due to low quality or extreme difficulty, we further refined our dataset by excluding images and referring expressions that did not meet the following criteria. For detailed examples of such exclusions, please refer to Fig. 1.

– **Target Size**: We excluded data where the size of the target object, as measured by the area of its bounding box, was larger than 10% of the total image area.
– **Image Ratio**: We excluded data with images whose width-to-height ratios were outside the range of 1.2-2.0 (based on a screen ratio of 1.6). We did this to eliminate very elongated images, which might distort normal viewing behavior.
– **Sentence Complexity**: We excluded data where the referring expression sentence was either too simple or too complex. We measured sentence complexity using the metric introduced by Liu et al. [13], which is correlated with sentence length and frequency. Specifically, we excluded data where the referring expression language complexity was below the 10th percentile (e.g., "the girl") or above the 90th percentile (e.g., "second row from left to right second one up from bottom..."). This ensured that the length of the remaining referring expressions ranged from 2 to 10 words, with a median of 4. The original dataset had a wider range of sentence lengths, from 1 to 39 words, with a median of 4.

After applying the aforementioned exclusion criteria, we were left with 7,568 referring expressions from 72 categories. However, this number exceeded our resources for gaze data collection. Therefore, we decided to further trim the dataset while maintaining a balanced distribution of target categories. To do so, we removed entire categories if the application of the exclusion criteria left fewer than 100 referring expressions per category. Then, we randomly selected up to 150 data points per category. This process resulted in a dataset of 2,422 referring expressions from 18 categories. Finally, we conducted a manual exclusion process to remove any referring expressions or images containing obscene, inappropriate, or irrelevant content (e.g., blood, nudity, slang). We also manually removed

any data points with spelling mistakes, or incorrect or poor target descriptions. In total, 328 data points were removed during this manual exclusion process, yielding 2,094 image-expression pairs for the dataset.

Fig. 1: Excluded Samples

2 Gaze Recording Methods

2.1 Participants

Our dataset was collected from 220 participants, consisting of 155 male, 63 female, and 2 non-binary individuals. Participants were undergraduate students from our institution who were recruited for extra credit in a psychology class and had normal or corrected-to-normal vision. The age range of participants was between 18 and 33 years. Among the participants, 28 were non-native English speakers but rated their fluency level as either very good or good. This study had Institutional Review Board (IRB) approval.

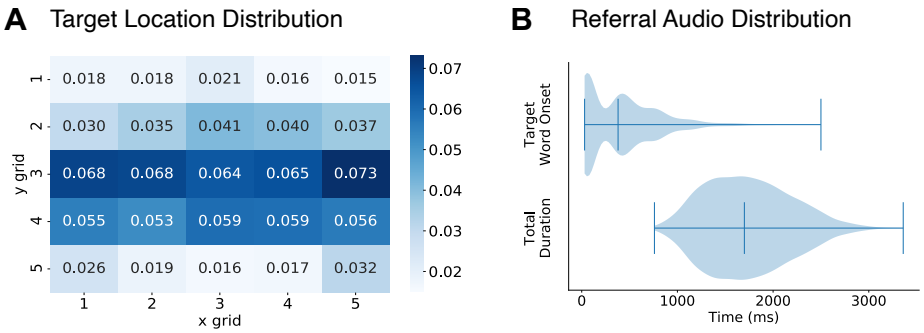

Fig. 2: Stimuli Statistics. A: Spatial distribution of target locations B: Temporal distribution of target word onset

2.2 Stimuli

Images were resized and padded to fit to the computer screen size and resolution (1680×1050 pixel resolution). Fig. 2A displays the spatial distribution of target locations in the images, which are evenly distributed but with slightly higher probability around the center-bottom area. Spoken referring expression were generated using Google Text to Speech API[5] commonly known as the gTTS available in Python. Fig. 2B shows the temporal distribution of target word onset, which was measured by the timing of the target word from generated audio. The original dataset did not provide the target word (i.e., the word in the sentence that refers to the target object). Therefore, we manually annotated the target word for each referring expression using the consensus of two annotators. Example target words are provided in Table 1. The target word is usually referred to at the beginning of the sentence (median time of 0.4 seconds), and the total duration of the audio mostly ranges from 1 to 3 seconds, with a median of 1.7 seconds.

[5] https://pypi.org/project/google-cloud-texttospeech/

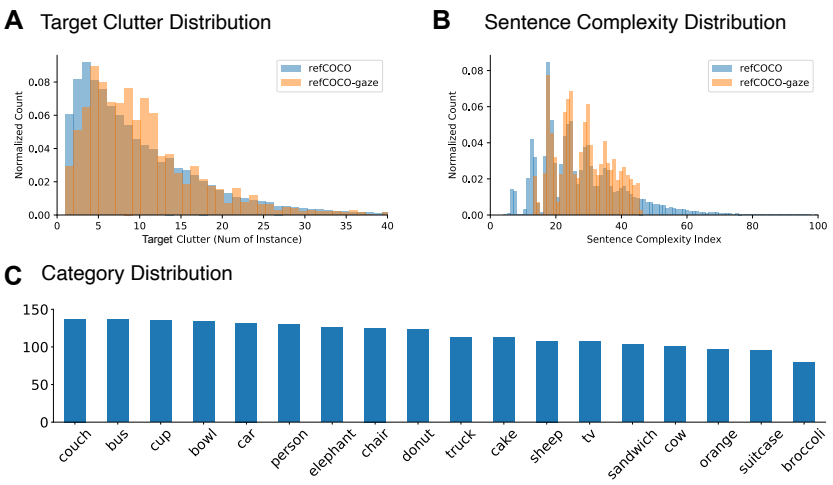

Fig. 3: Descriptive statistics for RefCOCO-Gaze images and referral expressions.

Our dataset closely matches the distribution of the original RefCOCO dataset (panels A and B of Fig. 3), capturing the distributions of target clutter (i.e., the number of instances of the same target category in the image) and sentence complexity (as measured by Liu *et al.* [13]). Fig. 3C shows the category distribution of target objects in the dataset. These categories were well-balanced and span a wide semantic range, from animate objects (e.g., person, sheep) and indoor objects (e.g., chair, cup) to outdoor vehicles (e.g., car, truck).

Table 1: Target Names

Category	Target Names Used for Referral
person	coach, figure, pic, jeans, player, girl, arm, row, thingy, shirt, girls, kid, hat, red, skier, picture, head, woman, left bottom, hand, slider, child, lady, man, boy, catcher, jacket, person, red and black, green, guy
elephant	rump and tail, ear, elephant butt, butt, one, camera, animal, elephant, baby, corner, legs
sheep	sheep, goat or sheep, cow, area, animal butt, animals face, leg, ship, one, lamb, sheep butt, animal, guy, face, goat, calf
cow	brown, cow, leg, camel, one, band, animal, bull, cows, corner, critter, legs, goat, calf
bus	trolley, double decker, van, phone, bus, ride, deck, decker, glass, train, truck, vehicle, thing, rectangle, car
car	taxi, benz, van, reflection, area, car, screen, mirror, police, suv, truck, vehicle, cab, ford, black
truck	hummer, area, van, semi, suv, door, car, vehicle, firetruck, bus, fedex, machine, thing, part, truck, item, fire truck, corner, rig, tarp, trailer
couch	loveseat, armchair, frame, seat, corner, ottoman, chair, love seat, thingy, couch, orange, seat cushion, pillows, leg rest, cushion, table, furniture, footstool, chairs, sofa, bed, thing, pillow
chair	woven, center, thing, chairs, tray, object, couch, jacket, item, bench, pattern, corner, chair, seat, lady, seat cushion
tv	laptop, monitor, tablet, tv, poster, screen, tv screen, bruce lee, face, computer, monitor screen, desktop, girl, sign, spot, computer screen
suitcase	case, box, container, briefcase, luggage, area, space, bag, black, item, suitcase, corner, chair, thing, trunk
bowl	right, bananas, cup and spoon, pot, dish, corner, thing, row, container, cup, things, section, bowl, butter, sauce, pan, kiwi, plate, food, left bowl, grapes, tuna, chips, broccoli, pottery piece, hot dog, fruit slices, soup, stuff, apples, dip
cup	tea, pot, whatever, one, juice, second, frosty, dish, candle, milk, blender, container, cup, section, jar, mug, beer, coffee, coke, drink, pitcher, toothbrush, glass, water, thing, stuff, bottle
donut	plate, sprinkles, item, food stuff, donuts, food, cheerio, donut, chocolate ice, bun, pastry, dessert, corner, skewer, doughnut, thing, striped, row
cake	muffin, pie, one, pastry, corner, thing, ice cream, row, item, pile, orange, hat, roll, plate, umbrella, food, dessert, brownie, cupcake, cake, fruit, cookie, frosting, chocolate, bread, center, biscuit, train car, cake slice
sandwich	taco, waffle, burger, ball, pastry, wrap, sammy, toast, flower, sub, bowl, palte, roll, plate, sandwich, food, appetizer, bun, half, banana slices, bread, meat, thing, piece
orange	one, slice, apple, left, thing, orange slice, bowl, oranges, pieces, lime, grapefruit, fruit, food, lemons, front, orange, lemon, stem, egg, row
broccoli	broccoli piece, broccoli pieces, greens, spinach, food, blur, veggie, patch, thing, piece, green, goop, basket, piece of broccoli, broccoli

2.3 Procedure and Apparatus

Gaze was recorded using the EyeLink 1000 eye-tracker (SR Research Ltd., Ottawa, Ontario, Canada) and the data were exported using the EyeLink Data Viewer software package (also from SR Research Ltd.). During the experiment, the presentation of images was controlled using Experiment Builder software (SR Research Ltd., Ottawa, Ontario, Canada). The stimuli were displayed on a 22-inch LCD monitor, positioned at a viewing distance of 47cm from the participant, with the help of chin and head rests. This resulted in a horizontal and vertical visual angle of $54^\circ \times 35^\circ$, respectively. At the beginning of each trial, participants were instructed to fixate on a central point but were free to move their eyes while searching for the target. Eye movements were recorded throughout the experiment using the EyeLink 1000 eye-tracker in tower-mount configuration. Prior to each block or whenever necessary, the eye-tracker was calibrated using a 9-point calibration method, and the calibration was not accepted unless the average calibration error was below 1.0° and the maximum error was below 1.5°. The experiment was conducted in a quiet laboratory room under dim lighting conditions. All responses were recorded using Microsoft Game controller triggers. The following instructions were provided to the participants prior to the gaze data collection process:

"We wish to observe your natural eye-movement behavior while searching for a referred target. You will be shown 100 images with spoken referring expressions describing the target's location and appearance. Your job is to find a target AS QUICKLY AND ACCURATELY AS POSSIBLE. When you find a target, please press any button on the top side of the controller. We will analyze your gaze later and measure accuracy by checking whether your gaze land on the target correctly at the time you press. So please make sure you press the button WHILE you are looking at the target. Please press the button as soon as you find the referred target. You can browse each image up to 5 seconds after the sound ends. There will be a break around halfway through the experiment, but if you need an additional break during the experiment, let the experimenter know anytime."

2.4 Preprocessing

Fixations were detected from raw gaze samples using the EyeLink online parser, which applied velocity and acceleration thresholds of $30^\circ/s$ and $8000^\circ/s^2$, respectively. Fixations with a duration lower than 60ms were filtered out, but all other fixations were retained. The initial raw dataset consisted of 21,898 valid scanpaths. However, to ensure data reliability, we removed trials where participants did not find the target within a given time limit of 5 seconds or reported not finding the target in the survey. We also eliminated trials where any of the participant's fixations did not land within the target bounding box, resulting in the removal of 10% of the entire dataset and leaving us with 19,738 scanpaths. Additionally, we observed that 6% of trials had the participant's final fixation not within the target area, which may have been due to them moving away from

the target as they pressed the button. To address this, we trimmed the fixations up to the last fixation that landed on the target, thereby ensuring that only fixations relevant to the target search were included in the analysis.

2.5 Gaze Data Analysis for RefCOCO-Gaze

RefCOCO-Gaze fixations are intention-driven. As can be seen by comparing the top two rows in Table 1 from the main text, the inter-observer agreement metrics (row 1, labeled "Human") far exceed the Random baseline metrics (row 2). Based on this observation, and findings of previous behavioral work [21, 29] suggesting that high inter-observer agreement mark similar task-driven attention allocation across individuals, we infer that the fixations in RefCOCO-Gaze are not random but rather, intention-driven and under attention control.

Target Localization analysis. Analysis of the gaze data collected in our incremental object referral task revealed that on 9.76% of the trials, participants failed to either fixate on the correct target or to localize the target within the 5 second limit. Mean saccade amplitudes for successful and failed localizations were 192.948 (standard deviation=75.86), and 191.662 (standard deviation=78.94), respectively. On the other hand, mean fixation durations (in msecs) for successful and failure localizations were 280.741 (standard deviation=114.38) and 277.432 (standard deviation=126.68), respectively. As is evident, the gaze statistics of average saccade amplitude and average fixation duration did not significantly differ between the successful localizations and failed localizations (T-test revealed p-value=0.437 for average saccade amplitude, and p-value=0.247 for average fixation duration – both not statistically significant since p-value>0.05). However failure cases yielded statistically significant longer scanpaths (average of 10 fixations for failure, 8 for success; p-value<0.05). Failure cases also showed strong positive correlations with scene complexity (Pearson correlation coefficient r=0.81) in terms of object instance count in a scene, and referral language perplexity (Pearson correlation coefficient r=0.76). These complexity scores tend to be higher for failure cases than for successful ones, suggesting that search performance decreases with increasing scene complexity and linguistic complexity of the referring expressions.

We also note that in 12.52% of the trials, observers fixated on the target during exposure time. Yet, for these trials, search ended after a median of 5 words, implying that observers required ample description for confident localization.

3 Comparison of RefCOCO-Gaze with other gaze datasets

Here, we compare related datasets discussed in Related Work section (Section 2 in the main text) in the table below. Our proposed dataset, RefCOCO-Gaze is the *only* gaze scanpath dataset for the incremental object referral task.

Dataset	Apparatus	Task	Gaze recorded [during/after] task description	Stimuli	No. of scanpaths	No. of Subjects	Relevant w.r.t. Object Referral	Relevant w.r.t. Incremental Prediction
COCO-Search18 [4]	Eye-tracker	Categorical Visual Search	After	Images	299037	10	No	No
AiR [2]	Eye-tracker	VQA	After	Images	13173	20	No	No
Localized Narratives [19]	Mouse proxy	Image Captioning	During	Images	848749	156	No	**Yes**
He *et al.* [10]	Eye-tracker	Image Captioning	During	Images	14000	16	No	**Yes**
SNAG [23, 24]	Eye-tracker	Image Description	During	Images	3000	30	No	**Yes**
OR [25]	Face videos	Object Referral	After	Videos	30000	20	**Yes**	No
Zhang *et al.* [33]	Gaze Following	Object Referral	After	Images	- *	-	**Yes**	No
RefCOCO-Gaze(ours)	Eye-tracker	Incremental Object Referral	During	Images	19738	220	**Yes**	**Yes**

* Zhang *et al.* [33] collect 40000 *static gaze heatmaps*, not *spatiotemporal gaze scanpaths*

4 Additional Details of ART

In this section, we share additional details about the ART model, such as details of implementation, architectural design, and hyperparameter choices for our experiments on RefCOCO-Gaze.

4.1 Visual Encoder and Language Encoder

For designing the visual encoder, we use an ImageNet [6] pre-trained ResNet-50 [9] backbone followed by a transformer encoder consisting of 6 standard transformer encoder layers [26] with hidden size $d_{vis} = 256$ and 8 attention heads. A dropout of 0.1 was applied to the transformer encoder layers. The output of the visual encoder is patch embedding tensor $g_{vis} \in \mathbb{R}^{d_{vis} \times hw}$, corresponding to $h \times w$ grid, where $h = 10, w = 18$. For the language encoder, we use the RoBERTa-base variant [14] which generates embeddings of dimension $d_{lang} = 768$ for each token in the tokenized text string. The hyperparameter l_{lang} is set to 32. RoBERTa encodes text tokenized using a Byte-Pair Encoding (BPE) [20]. RoBERTa is pre-trained on a large corpus of English data (which includes the BookCorpus [34], English Wikipedia data, the English portion of CommonCrawl News dataset [17] called CC-News, OpenWebText [8] and STORIES [22]) using a Masked Language Modeling (MLM) objective with a dynamic masking scheme. As mentioned in the main text, we specifically use ResNet-50 and RoBERTa backbones for fair comparison because they form the backbones of our baselines Chen *et al.* [3] and Gazeformer [16] variants. Both visual and language encoders are trainable, and not frozen as in Gazeformer [16].

4.2 Visuo-linguistic Transformer Encoder

For our experiments on RefCOCO-Gaze, ART's visuo-linguistic encoder consists of 6 standard transformer encoder layers [26] with hidden size (d) 256 and 8 attention heads each. A dropout of 0.1 was applied to all transformer layers in this module. For the bounding box regression and target category prediction heads, a dropout of 0.3 was applied during the pre-training phase while a dropout of 0.2 was applied during the training phase. To deal with scale variation, we normalize the parameters of of ground truth bounding boxes and consequently apply sigmoid activation to the bounding box regression head.

4.3 Pack Decoder & Fixation Prediction

For our experiments on RefCOCO-Gaze, ART's pack decoder module consists of 6 transformer decoder layers [26] with hidden size(d) 256 and 8 attention heads. A dropout of 0.2 was applied for all transformer decoder layers in this module. For the fixation prediction heads in the fixation prediction module, a dropout of 0.4 was applied. We choose hyperparameters $L_{\mathcal{P}}$ and $L_{\mathcal{C}}$ to be 6 and 36 respectively. Spatial location estimation was done by regressing parameters (i.e. mean and log-variance) of two separate Gaussian distributions using 4 regression heads (two heads each for the two Gaussian distrbutions - one head for estimating mean and the other head for estimating log-variance) in the fixation prediction module. These Gaussian distributions model the x and y co-ordinates (raw unnormalized pixel co-ordinates) of fixations [16]. The spatial locations are sampled from the Gaussian distributions using the reparameterization trick [11]. The range of the predicted unnormalized fixation location (x and y) co-ordinates are the respective

image dimensions. We do not involve the pack decoder module and the fixation prediction module in the pre-training stage.

4.4 Pre-training & Training

To deal with scale variation, we normalize the parameters of the ground truth bounding boxes during pre-training and training of the bounding box head. We use AdamW [15] optimizers for our pre-training and training phases with weight value 1e-4. During the pre-training process, the visual encoder, the language encoder and the visuo-linguistic encoder are all assigned learning rates of 1e-5. During the training process, the visual encoder and the language encoder are both assigned learning rates of 1e-7 while the visuolinguistic encoder is assigned a learning rate of 1e-5, while the rest of the ART model is assigned a learning rate of 1e-4. We pre-train on the RefCOCO training set for 200 epochs with a batch size of 128 and train on RefCOCO-Gaze training set with a batch size of 64 for a maximum of 200 epochs. Note that the visual, language and visuo-linguistic encoders are trainable (not frozen) during the pre-training stage, and all components of the ART model are trainable (not frozen) during the training stage. We ran our experiments on NVIDIA RTX A5000 GPUs.

5 Additional Details of Metrics

In this section, we provide additional implementation details of the scanpath metrics.

SS. This metric is the sequence score between the ground truth and predicted scanpaths over the *entire* referring expression. Hence, this metric considers *only* the valid 2D fixation locations.

SS$_{pack}$. We might encounter two edge cases while calculating SS_{pack} - either (1) the predicted pack is a null pack or (2) the ground truth pack is a null pack, with both scenarios resulting in empty strings which hinder direct application of string matching algorithm [18]. We handle the first scenario by duplicating the last fixation of previous non-null predicted pack (initial central fixation point in case there are no previous fixations) and handle the second scenario by duplicating the last fixation of previous non-null ground truth pack (initial central fixation point in case there are no previous fixations) - similar to the process for calculating ScanMatch with duration [5]. Similarly, we also duplicate the last 2D fixation when one of ground truth scanpath or predicted scanpath has terminated and the other one has not.

CC$_{pack}$. For our implementation of CC_{pack}, we add a small $\epsilon = 1e^{-9}$ to the ground truth and predicted maps to avoid a divide-by-zero error for cases where either the ground truth map or the prediction map is a zero map due to a null pack.

NSS$_{pack}$. We disregard cases where either ground truth or predicted pack is a null pack while calculating the average for NSS_{pack}. This is because there is no

Table 4: Performance of ART and baselines on RefCOCO-Gaze test set when trained and evaluated on both fixation location and fixation duration prediction tasks.

(a) Duration-agnostic metrics

	$SS \uparrow$	$SS_{pack} \uparrow$	$FED \downarrow$	$FED_{pack} \downarrow$	$CC_{pack} \uparrow$	$NSS_{pack} \uparrow$
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 [27]	0.216	0.170	17.084	2.901	0.174	2.175
Chen *et al.* [3]	0.281	0.255	6.825	1.163	0.209	1.953
Gazeformer-ref [16]	0.261	0.187	6.833	1.307	0.197	2.882
Gazeformer-cat [16]	0.244	0.172	7.144	1.394	0.194	2.664
ART (Proposed)	**0.356**	**0.285**	**6.410**	**1.161**	**0.281**	**3.539**

(b) Duration-aware metrics

	$SS^{(t)} \uparrow$	$SS^{(t)}_{pack} \uparrow$	$FED^{(t)} \downarrow$	$FED^{(t)}_{pack} \downarrow$	$MM_{dur} \uparrow$
Human	0.379	0.215	38.153	8.204	0.589
Random	0.169	0.097	108.296	18.395	0.688
OFA [27]	0.206	0.124	103.347	17.868	0.688
Chen *et al.* [3]	0.272	0.157	42.058	8.224	0.633
Gazeformer-ref [16]	0.236	0.166	39.104	7.131	0.617
Gazeformer-cat [16]	0.224	0.161	39.937	7.216	0.519
ART (Proposed)	**0.332**	**0.199**	**35.997**	**7.120**	**0.696**

theoretical upper or lower bound of NSS that can be assigned to scenarios where either one or both of ground truth and predicted packs are null packs (resulting in zero action/saliency maps).

6 Fixation Duration Prediction with ART

ART is also capable of predicting the fixation durations of humans. We model fixation durations as Gaussian distributions, similar to how we model fixation locations. First, we reparameterize a fixation $\mathbf{p}_i^k$ using five parameters: x-location x_i^k, y-location y_i^k, *fixation duration* d_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 then add two fixation duration regression heads (along with the already existing fixation location regression heads) to the fixation prediction module to estimate parameters (i.e., mean and log-variance) of a Gaussian distribution modeling fixation durations. Fixation durations d_i^k are sampled from this Gaussian distribution using the reparameterization trick [11]. Let the predicted pack of fixations $\mathcal{P}_k = \{(x_i^k, y_i^k, d_i^k)\}_{i=1}^{L_\mathcal{P}}$, and ground-truth pack of fixations $\hat{\mathcal{P}}_k = \{(\hat{x}_i^k, \hat{y}_i^k, \hat{d}_i^k)\}_{i=1}^{l^k}$ 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. For accomodating the additional fixation duration prediction objective, an L_1 loss $\mathcal{L}_d^k$ between ground-truth fixation durations $\hat{d}_i^k$ and predicted fixation durations d_i^k is added to the formulation of $\mathcal{L}_{gaze}$ in Equation 1 of the main paper. Hence, upon accounting for fixation duration prediction along with fixation location prediction, $\mathcal{L}_{gaze}$ for a minibatch of size M now becomes:

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

Here $\mathcal{L}_{xy}^k = \frac{1}{l^k}\sum_{i=1}^{l^k}\left(|x_i^k - \hat{x}_i^k| + |y_i^k - \hat{y}_i^k|\right)$, $\mathcal{L}_d^k = \frac{1}{l^k}\sum_{i=1}^{l^k}\left(|d_i^k - \hat{d}_i^k|\right)$, and $\mathcal{L}_{token}^k = -\sum_{i=1}^{L_\mathcal{P}}\sum_{t\in T}\hat{v}_{i,t}^k \log(v_{i,t}^k)$. Hence, the total multi-task loss $\mathcal{L}$ that we use to train our ART model for both fixation location prediction and fixation duration prediction 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. Note that $\mathcal{L}_{ground}$ is the auxiliary multi-task grounding loss defined in Sec. 4.2 of the main paper.

To evaluate fixation duration prediction of baselines and ART, we train them on fixation duration prediction along with fixation location prediction. Along with the duration-agnostic metrics we used in the main paper, we also report $SS^{(t)}$, $FED^{(t)}$, $SS_{pack}^{(t)}$, and $FED_{pack}^{(t)}$, which are duration-aware variants (as done in previous works [3, 5, 16]) of SS, FED, SS_{pack}, and FED_{pack}, respectively. We also report the duration component of MultiMatch [1, 7] (MM_{dur}). Higher SS, SS_{pack}, $SS^{(t)}$, $SS_{pack}^{(t)}$, CC_{pack}, NSS_{pack}, MM_{dur} metrics signify higher scanpath similarity, whereas higher FED, FED_{pack}, $FED_{pack}^{(t)}$, and $FED^{(t)}$ metrics denote lower scanpath similarity. The results are in Table 4.

ART outperforms baselines on all metrics (both duration-agnostic and duration-aware) when trained on and evaluated for fixation duration prediction and fixation location prediction. The model hyperparameters, pre-training and training processes remain as mentioned in Sec. 4. Additional details for baselines endowed with fixation prediction can be found in Sec. 8.

Note that MM_{dur} reflects solely the duration component, in contrast to the spatio-temporal metrics ($SS^{(t)}$, $FED^{(t)}$, $SS_{pack}^{(t)}$, $FED_{pack}^{(t)}$) in Table 4(b), and by that metric, the random baseline (whose generated fixation duration is set to the average training set fixation duration, as detailed in Sec. 8) scores higher than the human consistency score. We believe this is because of very poor agreement among the behavioral participants in their fixation duration in our incremental object referral task, which makes the prediction of fixation duration less meaningful than the prediction of fixation spatial locations (when we created scanpaths using average fixation locations from the RefCOCO-Gaze training set, we observed $SS = 0.037$ and $SS_{pack} = 0.044$, which are far lower than human consistency scores ($SS = 0.400$, $SS_{pack} = 0.317$), signifying that the spatial attention of humans for our task is meaningful and intention-driven).

Table 5: Performance of ART and baselines on COCO-Search18 [4] dataset. Gazeformer and ART models are shown in two variants - one with fixation prediction capability ("w/ dur." in parenthesis) and one without fixation prediction capability ("w/o dur." in parenthesis). Metrics in bold are the best performing metrics, while those underlined with a single dash are second-best, and those underlined with a double dash are third-best (we do not underline the third-best metric with double dash for duration-aware metrics since there are only three models predicting fixation duration).

(a) Duration-agnostic metrics

	$SS \uparrow$	$SemSS \uparrow$	$FED \downarrow$	$SemFED \downarrow$
Human	0.490	0.522	2.531	1.720
IRL [30]	0.405	0.441	2.781	2.393
Chen *et al.* [3]	0.398	0.425	2.376	2.064
FFM [31]	0.384	0.391	2.719	2.479
Gazeformer (w/o dur.) [16]	**0.475**	0.456	**2.159**	2.012
Gazeformer(w/ dur.) [16]	0.467	0.449	2.198	2.082
ART (w/o dur.) (Proposed)	0.454	**0.461**	2.251	**1.995**
ART(w/ dur.) (Proposed)	0.432	0.441	2.335	2.070

(b) Duration-aware metrics

	$SS^{(t)} \uparrow$	$SemSS^{(t)} \uparrow$	$FED^{(t)} \downarrow$	$SemFED^{(t)} \downarrow$	$MM_{dur} \uparrow$
Human	0.409	0.433	11.526	8.389	0.663
Chen *et al.* [3]	0.354	0.368	11.610	9.991	0.691
Gazeformer(w/ dur.) [16]	**0.417**	**0.408**	**10.216**	**8.771**	**0.727**
ART (w/ dur.)	0.373	0.394	11.127	9.089	0.725

7 ART generalizes to Categorical Search (COCO-Search18)

In this section, we extend ART to the related categorical search task. We do this via providing a prefix in the form of the category name (*e.g.*, "car" or "potted plant") to ART. We chose the large-scale categorical search fixation prediction dataset, COCO-Search18 [4] to train and evaluate ART and other baselines on its target-present trials. We use several competitive baselines, such as IRL [30] and FFM [31] which are not trained on an additional fixation duration prediction objective, along with Chen *et al.* [3]'s model and Gazeformer [16] which can be trained on the additional fixation duration prediction objective. State-of-the-art baseline Gazeformer [16] and ART are trained and evaluated in two variants - one which is trained on the additional fixation duration prediction objective ("w/ dur." in parenthesis) and another one which is *not* trained on the additional fixation duration prediction objective ("w/o dur." in parenthesis).

To evaluate on this categorical search task embodied by COCO-Search18, we follow previous methods [16, 31] and map all predictions to our input grid, and

then report Sequence Score [30] (SS), Semantic Sequence Score [31] ($SemSS$), Fixation Edit Distance [16] (FED), Semantic Fixation Edit Distance [16] ($SemFED$) and duration component of MultiMatch [1,7] (MM_{dur}). $SemSS$ and $SemFED$ differs from SS and FED, respectively, in that they convert scanpaths to strings of fixated scene object IDs instead of cluster IDs. We also report SS, $SemSS$, FED, and $SemFED$ with duration denoted by $SS^{(t)}$, $SemSS^{(t)}$, $FED^{(t)}$, and $SemFED^{(t)}$, respectively, as done by previous works [3, 5, 16]. Results are in Table 5. Higher SS, $SemSS$, $SS^{(t)}$, $SemSS^{(t)}$, MM_{dur} metrics signify higher scanpath similarity, whereas higher FED, $SemFED$, $FED^{(t)}$, and $SemFED^{(t)}$ metrics denote lower scanpath similarity.

Even though ART is designed for the incremental object referral task, in an extension to the categorical search task, we found that its performance is on par with Gazeformer [16], the state-of-the-art search fixation prediction model. ART even outperforms Gazeformer on Semantic Sequence Score (SemSS) and Semantic Fixation Edit Distance (SemFED) metrics. This generalization to the categorical search task further demonstrates the strength of ART's architecture.

8 Additional Details of Baseline Models

In this section, we provide additional implementation details of the baselines used in the main paper.

Random Scanpath: We sample pack length l_p uniformly from integers [0,1,...,$L_{\mathcal{P}}$] where $L_{\mathcal{P}}$ is the hyperparameter for maximum number of fixations in a pack. Since we chose $L_{\mathcal{P}} = 6$ for ART, we use the same value for this baseline for fair comparison. Then we uniformly sample l_p fixation locations within the entire image to obtain a generated pack of fixations. For the variant with fixation duration (Sec. 6), we use the average of all fixation durations in the RefCOCO-Gaze training set.

OFA: We sample pack length l_p uniformly from integers [0,1,...,$L_{\mathcal{P}}$] where $L_{\mathcal{P}}$ is the hyperparameter for maximum number of fixations in a pack. Since we chose $L_{\mathcal{P}} = 6$ for ART, we use the same value for this baseline for fair comparison. In order to obtain a generated pack of fixations, we uniformly sample l_p fixation locations from within the bounding box predicted by the OFA [27] model for the referring expression prefix corresponding to an incoming word within the referring expression. For the variant with fixation duration (Sec. 6), we use the average of all fixation durations in the RefCOCO-Gaze training set.

Chen *et al.* We train model from Chen *et al.* [3] using teacher-forcing algorithm [28] in the same manner we have trained ART. To incorporate fixation history, we construct a composite action map containing all previous fixations. We subsequently initialize the dynamic memory of the model with the sum of the task guidance map (from MDETR model which is pre-trained on RefCOCO for fair comparison) and the composite action map. Maximum number of fixations in a predicted pack is set to 6, identical to the value of pack length $L_{\mathcal{P}}$ chosen for ART, for fair comparison. In the context of experiments on RefCOCO-Gaze,

we train Chen *et al.*'s model with fixation duration information for results in Sec. 6 and without fixation duration information for the rest, unless specified otherwise.

Gazeformer-ref. We train the Gazeformer [16] variant, that we named Gazeformer-ref, using teacher-forcing algorithm [28] in the same manner we have trained ART. We provide previous fixation information in the form of the last fixation from the previous non-null pack, which is encoded using a 2D positional encoding and added to the first fixation query as prescribed in [16] for including initial fixation information. The validity prediction head in the model is also extended to support the prediction of an additional end-of-scanpath token. Maximum number of fixations in a predicted pack is set to 6, identical to the value of pack length $L_{\mathcal{P}}$ chosen for ART, for fair comparison. In the context of experiments on RefCOCO-Gaze, we train Gazeformer-ref with fixation duration information for results in Sec. 6 and without fixation duration information for the rest, unless specified otherwise.

Gazeformer-cat. We train the Gazeformer [16] variant, that we named Gazeformer-cat, using teacher-forcing algorithm [28] in the same manner we have trained ART. The previous fixation history is conveyed in the same manner as in the implementation of Gazeformer-ref (see above). We also extend the validity prediction head to support the prediction of an additional end-of-scanpath token. The target category estimation which is used to construct the input category name comes from a RoBERTa-based classifier which is separately trained on RefCOCO referring expressions and their corresponding target categories. Specifically, the target category estimator is a RoBERTa-base model with a classification head on top. This baseline should show how important target category estimation is for gaze prediction. Maximum number of fixations in a predicted pack is set to 6, identical to the value of pack length $L_{\mathcal{P}}$ chosen for ART, for fair comparison. In the context of experiments on RefCOCO-Gaze, we train Gazeformer-cat with fixation duration information for results in Sec. 6 and without fixation duration information for the rest, unless specified otherwise.

9 Additional metrics for Ablation Studies

In Table 6, we augment Table 2 in the main paper with additional metrics (FED, FED_{pack} and NSS_{pack}). The trends remain similar to the what we observed for SS and SS_{pack} scores, thereby reaffirming our assertion that *both* object localization and target category prediction tasks are integral to the object referral process and that pre-training on these tasks is instrumental for superior performance.

10 Additional Ablation Studies and Analysis

We provide two additional ablations (in addition to the five ablations in Table 2 of the main paper and Table 6 in Section 9) tabulated as ablation #1 and ablation #2 in Table 7. As it can be seen, addition of only one of the auxiliary losses

Table 6: Ablation studies on ART model (reported in Table 2 of the main paper) augmented with additional metrics. 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.

Ablation #	Pre-training	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	SS ↑	SS_{pack} ↑	FED ↓	FED_{pack} ↓	CC_{pack} ↑	NSS_{pack} ↑
1	×	×	×	0.309	0.257	6.873	1.203	0.222	3.032
2	✓	✓	×	0.321	0.279	7.341	1.348	0.239	2.769
3	✓	×	✓	0.292	0.260	6.713	1.162	0.216	2.967
4	×	✓	✓	0.304	0.257	7.104	1.245	0.215	2.953
5	✓	✓	✓	**0.359**	**0.292**	**6.371**	**1.143**	**0.280**	**3.478**

Table 7: Additional 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. Ablations #3, #4 and #5 are from Table 6 (also in Table 2 in the main paper).

Ablation #	Pre-training	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	SS ↑	SS_{pack} ↑	FED ↓	FED_{pack} ↓	CC_{pack} ↑	NSS_{pack} ↑
1	×	✓	×	0.319	0.270	6.736	1.193	0.228	3.135
2	×	×	✓	0.309	0.264	6.705	1.175	0.216	2.919
3	✓	✓	×	0.321	0.279	7.341	1.348	0.239	2.769
4	✓	×	✓	0.292	0.260	6.713	1.162	0.216	2.967
5	✓	✓	✓	**0.359**	**0.292**	**6.371**	**1.143**	**0.280**	**3.478**

$\mathcal{L}_{bbox}$ and $\mathcal{L}_{target}$ results in little to no boost in performance. It is evident that we need *both* $\mathcal{L}_{bbox}$ and $\mathcal{L}_{target}$ in the objective function *along with pre-training* for our model to achieve high performance. We posit that ablation #4 in Table 7 fails to perform well because the target category prediction task largely, if not completely, relies on the linguistic input (i.e., the referring expression). Consequently, the sub-networks dedicated to visual and visuo-linguistic processing (especially, the visual encoder) might not benefit from pre-training only on this objective whereas the linguistic subnetworks (i.e. the linguistic encoder) are greatly optimized. We speculate that it is also perhaps hard for ART to adapt to object referral during training after pre-training its parameters to significantly align with the target-category estimation objective (which can be inadequate for our task since there are multiple objects belonging to the target category) in ablation #4 of Table 7. On the other hand, object localization seems to be much more aligned with the object referral task, which is indeed shown by ablation #3 in Table 7. In summary, the ablation studies validate our hypothesis that *both* object localization and target category prediction tasks are integral to the object referral process. Also note that Ablations 1 and 4 in Table 6 and Ablations 1 and 2 in Table 7 show that even when ART is not pre-trained on RefCOCO,

Table 8: Additional ablation studies on ART model when trained with additional Fixation Duration Prediction objective. 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.

(a) Duration-agnostic Metrics

Ablation #	Pre-training	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	SS ↑	SS_{pack} ↑	FED ↓	FED_{pack} ↓	CC_{pack} ↑	NSS_{pack} ↑
1	×	×	×	0.206	0.169	10.840	2.073	0.167	2.331
2	×	✓	×	0.262	0.230	8.225	1.528	0.207	2.930
3	×	×	✓	0.269	0.201	9.626	1.616	0.196	2.396
4	×	✓	✓	0.296	0.252	7.049	1.271	0.209	2.914
5	✓	✓	×	0.309	0.278	7.306	1.339	0.257	3.307
6	✓	×	✓	0.284	0.210	7.172	1.351	0.179	2.460
7	✓	✓	✓	**0.356**	**0.285**	**6.410**	**1.161**	**0.281**	**3.539**

(b) Duration-aware Metrics

Ablation #	Pre-training	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	$SS^{(t)}$ ↑	$SS^{(t)}_{pack}$ ↑	$FED^{(t)}$ ↓	$FED^{(t)}_{pack}$ ↓	MM_{dur} ↑
1	×	×	×	0.222	0.110	55.614	11.274	0.672
2	×	✓	×	0.253	0.169	44.074	8.559	0.691
3	×	×	✓	0.277	0.159	47.876	8.928	0.675
4	×	✓	✓	0.253	0.164	38.722	7.232	0.652
5	✓	✓	×	0.282	0.181	37.922	7.389	0.685
6	✓	×	✓	0.252	0.165	38.934	7.206	0.683
7	✓	✓	✓	**0.332**	**0.199**	**35.997**	**7.120**	**0.696**

it still outperforms baselines that are also not pre-trained on RefCOCO, i.e., Gazeformer-ref and Gazeformer-cat (in Table 1 of main text).

In Table 8, we tabulate the ablation studies for ART when equipped with fixation duration prediction capability. As shown in Sec. 6 through analysis of MM_{dur} metric values for random baseline and human consistency, there is poor agreement between participants in their fixation durations and thus training on such noisy supervision can be challenging. We interpret Table 8 as being consistent with our findings from the ablation studies with ART without fixation prediction (Sec. 5.4 in main text; Sec. 9, Sec. 10 in the supplement) in supporting our assertion that *both* pre-training and training on *both* auxiliary object localization and target category estimation objectives are crucial for ART's performance. We also observe that without pre-training and training on auxiliary losses, ART struggles to generalize when trained with noisy fixation durations, as seen in Ablation 1. Our proposed model (#7 in Table 8) supports our assertion that ART's pre-training and training on the two object grounding tasks for complete/partial expressions underlying object referral significantly contributes towards its SOTA performance when compared to existing baselines that are

unable to pre-train/train on object grounding for partial/complete expressions. Our proposed model thus generalizes well even when trained with a noisy supervision signal, such as the fixation durations. We also note that when not pre-trained on RefCOCO (Ablations 2, 3, and 4 in Table 8), ART still outperforms baselines that are not pre-trained on RefCOCO, i.e. Gazeformer-ref and Gazeformer-cat (see Table 4).

11 Auxiliary Next Word Token Prediction Task

We also hypothesized that predicting the next linguistic token during search also underlies incremental object referral process along with the object localization and target category prediction tasks. So we added a `NEXT_WORD_TOKEN` token along with `BBOX` and `TGT` tokens as input to the visuo-linguistic encoder. The corresponding latent vector served as input to an MLP which generated logits over a vocabulary of tokens in order to predict the next word token. The loss imposed is a cross-entropy loss $\mathcal{L}_{nextword}$. The results are tabulated in Table 9. As we can see, adding $\mathcal{L}_{nextword}$ does not improve the performance significantly – we achieve best performance without $\mathcal{L}_{nextword}$ (Ablation #1 in Table 9). We hypothesize that this is because the next word token prediction task is considerably more difficult than the object localization and target category prediction tasks, and potentially introduces noise while training on the gaze prediction objective.

Table 9: Effect of auxiliary next word token prediction task on ART model. Ablations #1 and #3 are from Table 2 in main text.

Sl.No#	Pre-training	$\mathcal{L}_{nextword}$	$\mathcal{L}_{bbox}$	$\mathcal{L}_{target}$	SS	SS_{pack}	CC_{pack}	NSS_{pack}
1	✓	×	✓	✓	**0.359**	**0.292**	**0.280**	**3.478**
2	✓	✓	✓	✓	0.355	0.281	0.269	3.388
3	×	×	✓	✓	0.304	0.257	0.215	2.953
4	×	✓	✓	✓	0.313	0.265	0.224	3.049

12 Qualitative Results

In this section, we present additional qualitative results of human behavior, our model ART and other competitive baseline models in Fig. 4 and Fig. 5. We see that ART efficiently finds the correct target through scanpaths that closely resemble human behavior in all rows except the last row in Fig. 5 - where it fails to localize the "head" of the correct person. In the first row of Fig. 4, we can see both the human participant and ART *wait* until the last word "left" to localize the correct muffin. We see a similar *waiting* pattern in second and third row of Fig. 4 where ART and the human participant wait until disambiguation towards

the end of the expression for localizing the correct "bus" and "car" respectively. In the first row of Fig. 5, we see a *scanning* behavior where ART and the human fixate on the kids in the center after hearing the word "kid" until the contextual information "far right" is provided in the end to locate the correct "kid". In the second row of Fig. 5, we observe that both ART and the human participant *wait* till the utterance of the target category word "elephant" in order to localize the correct elephant.

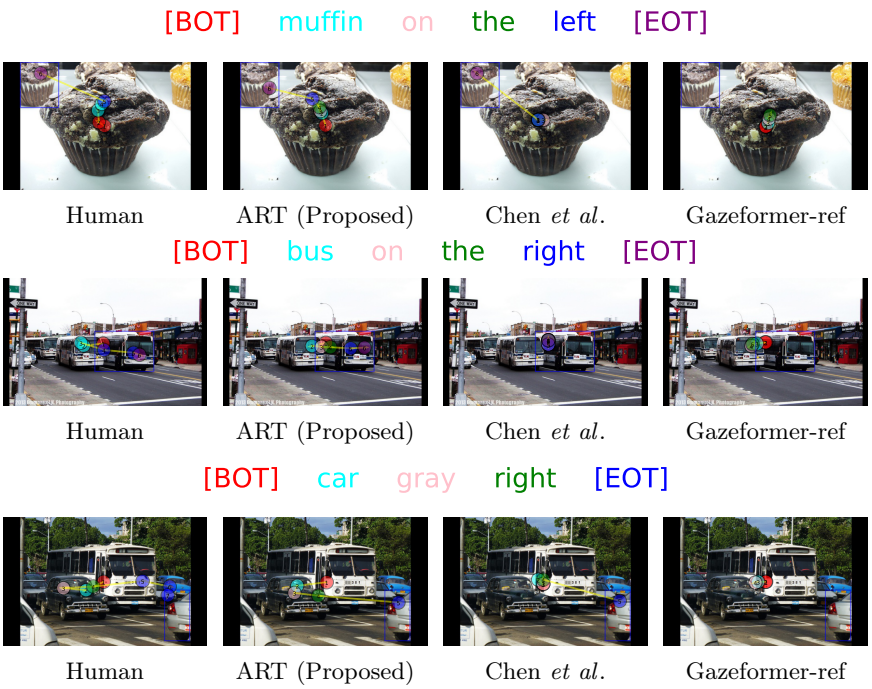

Fig. 4: Qualitative results [1/2]. Scanpaths from humans and three scanpath prediction models on three trials exhibiting strategic fixation behavior. Fixations (denoted by circles numbered with fixation order) are color-coded to corresponding words in the referring expression (above each row). Fixations 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.

Fig. 5: Qualitative results [2/2]. 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.

References

1. Anderson, N.C., Anderson, F., Kingstone, A., Bischof, W.F.: A comparison of scanpath comparison methods. Behavior research methods **47**(4), 1377–1392 (2015)
2. Chen, S., Jiang, M., Yang, J., Zhao, Q.: Air: Attention with reasoning capability. In: European Conference on Computer Vision (2020)
3. 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)
4. 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)
5. Cristino, F., Mathôt, S., Theeuwes, J., Gilchrist, I.D.: Scanmatch: A novel method for comparing fixation sequences. Behavior research methods **42**(3), 692–700 (2010)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2009)
7. Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., Holmqvist, K.: 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)
8. Gokaslan, A., Cohen, V.: Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus` (2019)
9. 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)
10. 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)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. 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)
13. Liu, X., Lai, H., Wong, D.F., Chao, L.S.: Norm-based curriculum learning for neural machine translation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020)
14. 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)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Mondal, S., Yang, Z., Ahn, S., Samaras, D., Zelinsky, G., Hoai, M.: Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
17. Nagel, S.: Cc-news. `http://commoncrawl.org/blog/news-dataset-available` (2016)
18. 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)

19. 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)
20. 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)
21. Torralba, A., Oliva, A., Castelhano, M.S., Henderson, J.M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review **113**(4), 766 (2006)
22. Trinh, T.H., Le, Q.V.: A simple method for commonsense reasoning. arXiv preprint arXiv:1806.02847 (2018)
23. 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)
24. 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)
25. 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)
26. 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)
27. 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)
28. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation **1**(2), 270–280 (1989)
29. Xu, J., Yue, S., Menchinelli, F., Guo, K.: What has been missed for predicting human attention in viewing driving clips? PeerJ **5**, e2946 (2017)
30. 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)
31. Yang, Z., Mondal, S., Ahn, S., Zelinsky, G., Hoai, M., Samaras, D.: Target-absent human attention. In: European Conference on Computer Vision (2022)
32. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: European Conference on Computer Vision (2016)
33. 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)
34. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2015)