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

ECCV 2024

Incremental object referral task

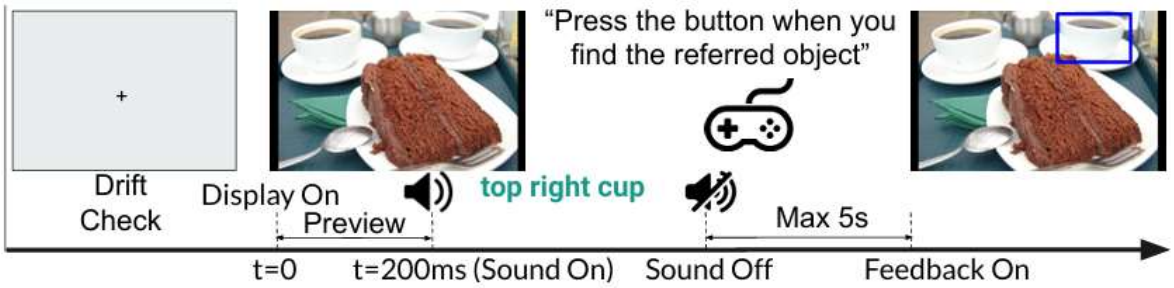
Scanpath prediction is a dynamic process of predicting human attention.

Incremental object referral task crementally predict eye movements of humans searching for a target object in an image as they are hearing a referring expression describing that target.

Contributions:

(1) Introducing the **incremental object referral task for gaze prediction** that will lead to more user-responsive HCI sys tems.
(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.

Datasets(Refcoco-Gaze)



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 images and listened to the associated referring expressions from the RefCOCO dataset. RefCOCO was collected using the ReferItGame where players must construct efficient referring expressions for another player to locate the correct object.

Datasets(Refcoco-Gaze)

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.

Architecture(ART)

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.

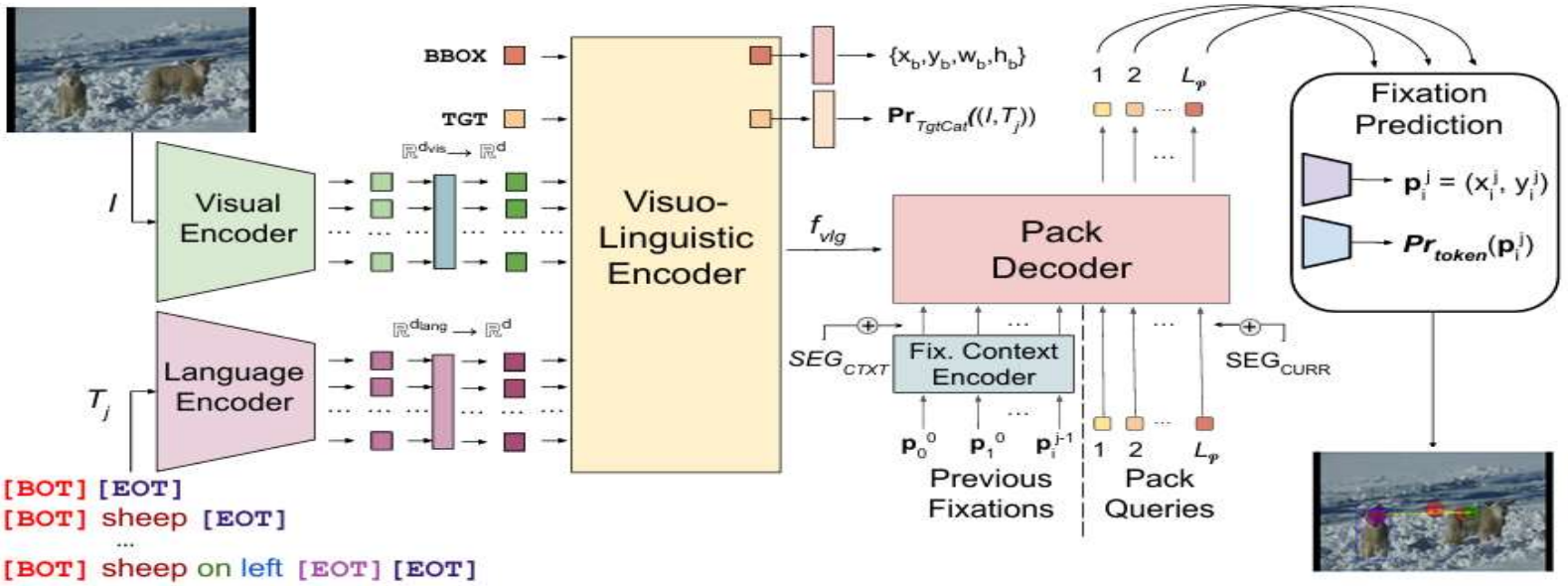


Fig. 3: Attention in Referral Transformer (ART) Architecture. On each pass after comprehending a new word, the model takes an image I and tokens T_j of prefix R_j of the referring expression as input and generates a possibly empty sub-sequence of fixations based on previous fixation history encoded by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

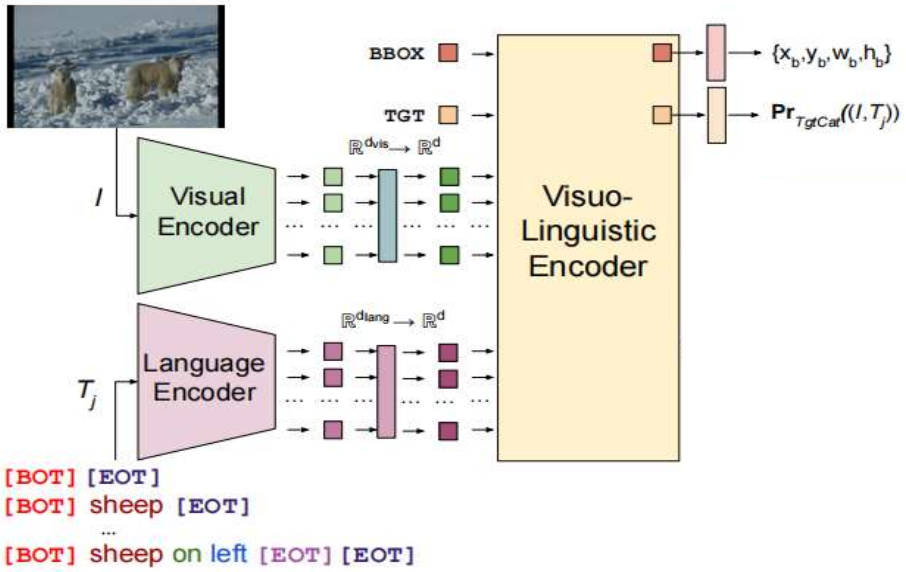
Pre-training

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 training data.

We apply an L_1 regression loss L_{reg} and a generalized IoU (GIoU) loss L_{giou}, between predicted and ground truth bounding box parameters. Object type loss L_{target} from the expression.

$$\mathcal{L}_{pretrain} = \mathcal{L}_{reg} + \mathcal{L}_{giou} + \mathcal{L}_{target}$$

$$\mathcal{L}_{bbox}$$



Training

To train ART on the gaze prediction task, we apply L1 regression loss on the predicted x and y locations.

For each of the L_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.

The multitask loss for a minibatch of size M is :

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

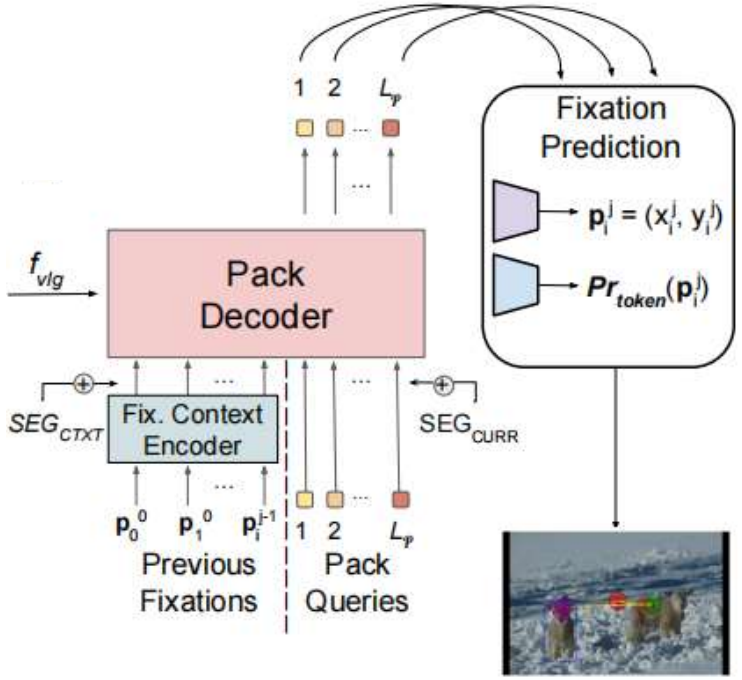
Where:

$$\mathcal{L}_{xy}^{k} = \frac{1}{l^k}\sum_{i=1}^{l^k}\left(|x_i^k - \hat{x}_i^k| + |y_i^k - \hat{y}_i^k|\right)$$

$$\mathcal{L}_{token}^{k} = -\sum_{i=1}^{L_{\mathcal{P}}}\sum_{t\in T}\hat{v}_{i,t}^{k}\log(v_{i,t}^{k})$$

Inference

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.



Quantitative comparison

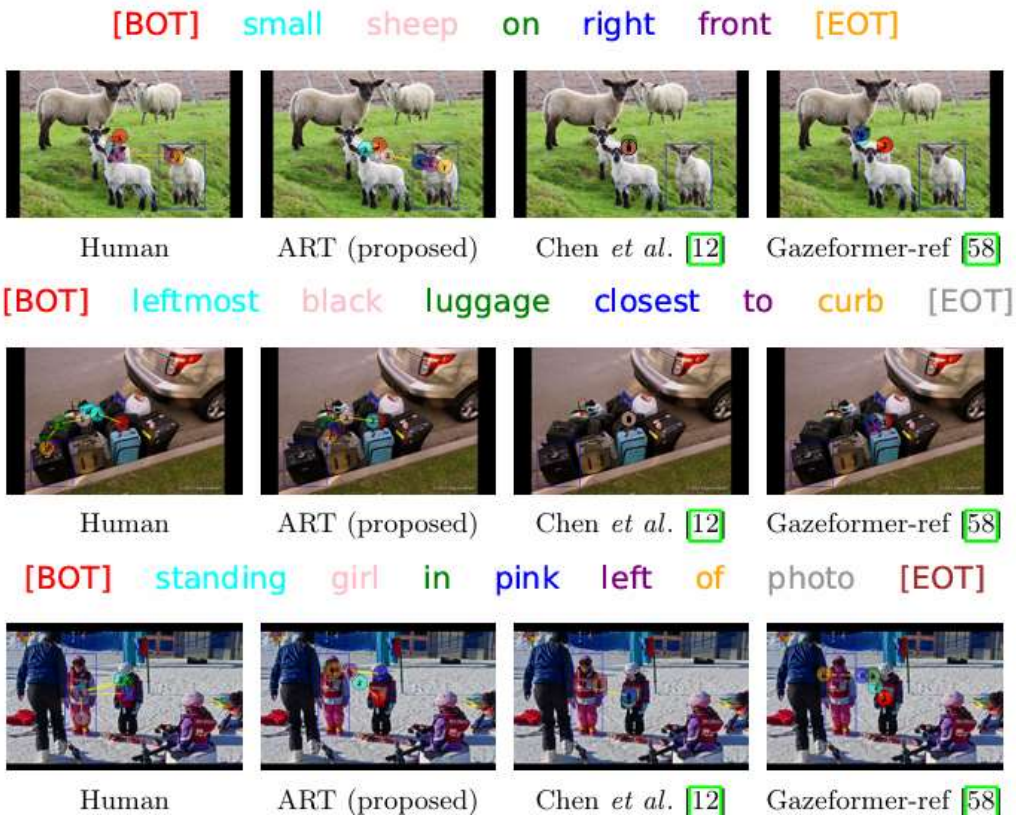
Table 1: Performance of ART and baselines on RefCOCO-Gaze test set.

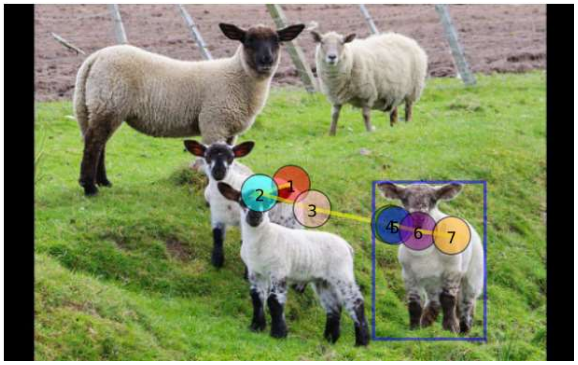
	$SS \uparrow$	$SS_{pack} \uparrow$	$FED \downarrow$	$FED_{pack} \downarrow$	$CC_{pack} \uparrow$	$NSS_{pack} \uparrow$
Human	0.400	0.317	6.573	1.278	0.283	3.112
Random	0.189	0.133	17.735	3.005	0.094	1.689
OFA [78]	0.216	0.170	17.084	2.901	0.174	2.175
Chen *et al.* [12]	0.299	0.188	8.309	1.507	0.159	1.557
Gazeformer-ref [58]	0.269	0.194	6.788	1.286	0.208	3.006
Gazeformer-cat [58]	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**

Qualitative comparison

ART also exhibits several strategic fixation patterns that we observe in the human data. (waiting, verification, and scanning)

1. For example, In the top row, ART waits near the center until after getting the word “right”, which conveys information about the referred sheep.
2. 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)





Ablation

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.

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

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**