Learning to Refer from Estimated Listener Gaze

Téa Wright & Alane Suhr
Department of Computer Science
University of California, Berkeley
Berkeley, CA 94704, USA
{teaywright, suhr}@berkeley.edu

Abstract

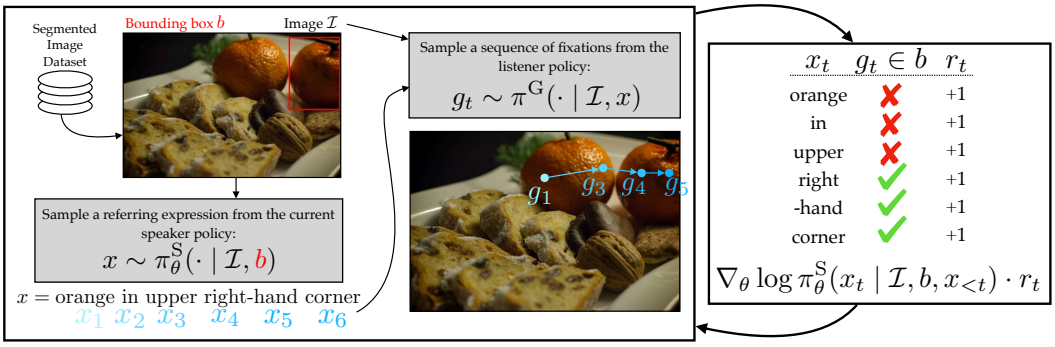
We propose to finetune vision-language models to generate more pragmatically optimal referring expressions by transforming observations of incremental listener comprehension, in the form of gaze scanpaths, into learning signals. During training, referring expressions are sampled from the speaker policy being optimized, conditioned on images and target referents; then, a neural listener estimating human gaze behavior maps from images and sampled referring expressions to scanpaths, each represented by a sequence of fixations, with each fixation corresponding to a word in the referring expression. We experiment with several approaches to convert fixation sequences and target referents into token- and sequence-level rewards, which are used to optimize policy parameters. Through evaluation with human listeners, we find that speaker policies trained with gaze-estimating listeners result in significantly more pragmatically-optimal references than base models, reducing sequence length from 15.4 down to 4.0 words while increasing referential success from 75.2 up to 80.0%. Our work demonstrates a promising opportunity for learning to generate utterances through language-based interaction, not only from the explicit signal of communicative success, but also from implicitly-available observations of a listener's process of comprehension. [1]

1 Introduction

Language-based interaction is collaborative. People collaborate through language use to come to a shared understanding of the world, and they do so efficiently by minimizing the total combined effort necessary to reach mutual understanding (Clark & Wilkes-Gibbs, 1986). In natural human interaction, we minimize joint effort through communication signals beyond the textual content of our utterances, including backchanneling, paralanguage, and gaze. Reference games are an ideal testbed to understand and measure these phenomena. Here, a speaker and a listener collaborate to establish joint attention to a part of their shared environment (Dale & Reiter, 1995). The speaker's task, known as referring expression generation (REG), is to produce an utterance that uniquely identifies the region of interest to their listener. The listener's task is to identify the target based on the speaker's description, called referring expression comprehension (REC). In reference games and elsewhere, people design their utterances by taking into account the shared interaction situation and their knowledge of their audience so that their contributions to the interaction are as informative as necessary, but not overly verbose (Grice, 1975; Kapur et al., 2026).

On the other hand, neural language models do not exhibit this level of pragmatic optimality. Ma et al. (2025) finds that referring expressions generated by modern vision-language models (VLMs) in reference games differ greatly from human-written references; they tend to include unnecessary information and often fail to uniquely identify referents by taking into account the interaction context. Recent work has approached the problem of improving referential success through learning from interaction, where a VLM policy is rewarded for

[1] Code and model weights are available at: https://github.com/Berkeley-NLP/reg-from-gaze



(a) Interaction **(b) Policy Update**

Figure 1: Learning to refer from interaction with incremental listeners. During training, we iterate between (a) deploying a speaker policy π^{S}_θ in a reference game with a listener policy π^{G}, which is trained to estimate gaze behavior of human listeners, and (b) updating the speaker policy by transforming observations of listener gaze behavior to rewards.

successfully describing an intended referent to a listener (Gul & Artzi, 2024; Tang et al., 2024). These policies are trained only from the outcome of reference games. Instead, Jara-Ettinger & Rubio-Fernandez (2022) argue that REG is best understood as a pragmatic effort to guide the listener's visual search process, observations of which are used by speakers to check listener understanding during language production (Clark & Krych, 2004).

We use observations of the process of visual search during reference comprehension as a learning signal for improving VLMs as speakers in reference games. Gaze provides a quantifiable signal that allows us to directly relate a speaker's language to a listener's visual search process; we can test exactly which tokens meaningfully contribute to the listener's understanding. To this end, we design a learning setup where VLMs are finetuned as speakers in interaction with gaze-estimating listeners. First, we train a VLM to estimate listener gaze. Then, during training of the VLM speaker, we iterate between deploying it in reference games with the gaze-estimating listener (Figure 1a), and optimizing the speaker policy using rewards derived from listener gaze estimates (Figure 1b).

We experiment with a variety of reward functions for converting gaze estimates to learning signals, and compare our approach to learning from non-gaze-estimating listeners. Our results show that using gaze-estimating listeners during training can both boost referential success in reference games with human listeners from 75.2 up to 80.0%, while also reducing the verbosity of such references from 15.4 down to 4.0 words on average. Our results also reveal a clear tradeoff in current VLMs between accuracy and verbosity: while speakers trained with non-gaze-estimating listeners achieve even higher referential success (up to 86%), this comes at a cost of generating increasingly longer references, up to an average of 24.4 words. Overall, we find that learning from estimated observations of the *process* of reference comprehension, rather than the *outcome* of a reference game alone, results in speaker policies that generate much more pragmatically optimal references.

2 Technical overview

Problem statement Let $x = \langle x_1, \ldots, x_T \rangle \in \mathcal{V}^+$ be a natural language referring expression comprised of tokens from a vocabulary $\mathcal{V}$, $\mathcal{I} \in \{0, \ldots, 255\}^{M \times N \times 3}$ an image of size M by N, and $b = (b_x, b_y, b_w, b_h) \in \mathbb{Z}^4$ a bounding box circumscribing a target referent.[2] We study the problem of referring expression generation, where a speaker policy π^{S}_θ maps from an image $\mathcal{I}$ and target bounding box b to a referring expression x. The task of referring expression

[2] $0 \leq b_x \leq M$ and $0 \leq b_y \leq N$ represent the anchor point of the bounding box at its top left corner; $0 < b_w \leq M - b_x$ and $0 < b_h \leq N - b_y$ represent the width and height of the bounding box.

comprehension (REC) is executed by a listener policy π^{REC}, which maps from the image $\mathcal{I}$ and the generated referring expression x to a coordinate $r = (r_x, r_y) \in \{0, \ldots, M\} \times \{0, \ldots, N\}$. Communicative success is achieved if the listener correctly identifies the target referent circumscribed by b; i.e., $r \in b$.[3] Success requires that x describes the target referent accurately and unambiguously with respect to the entire image. For every image and target referent, there are a large number of possible referring expressions that satisfy these criteria; one open challenge, then, is achieving referential efficiency by identifying the shortest such referring expression. This is the focus of our study. We extend this classical formulation of referring expression comprehension, where a listener maps from an entire referring expression to a single candidate coordinate, to an *incremental* comprehension task. Let a scanpath $g = \langle g_1, \ldots, g_T \rangle$ be a sequence of gaze fixations, with each g_i taking on either a null value $\emptyset$, or a coordinate value $\left(g_x^{(i)}, g_y^{(i)}\right) \in \{0, \ldots, M\} \times \{0, \ldots, N\}$. Here, a listener policy π^{G} maps from an image $\mathcal{I}$, (possibly partial) referring expression x of length $t \leq T$, and a scanpath prefix g of length $t-1$, to the next fixation g_t. Reference comprehension is performed iteratively; each fixation $g_t \sim \pi^{\text{G}}(\cdot \mid \mathcal{I}, x_{1:t}, g_{1:t-1})$ corresponds to an input word x_t.

Approach Our goal is to learn a speaker policy π_θ^{S} that generates referring expressions which are both *effective* and *efficient*. π^{S} is instantiated as a large VLM, with initial parameters θ_0 equivalent to the pretrained Molmo-7B-D model (Deitke et al., 2024). $\pi_{\theta_0}^{\text{S}}$ is neither efficient nor effective as a speaker; references generated from this policy are overly verbose and result in low communicative success. We finetune this policy through reinforcement learning to maximize reward (Section 5). During training, we simulate REC with a listener model π_ϕ^{G} that we have trained to estimate human gaze scanpaths (Section 4). Speaker reward is a function not only of final communicative success, but also of the incremental process of reference comprehension as estimated by the listener model $g \sim \pi_\phi^{\text{G}}(\cdot \mid \mathcal{I}, x)$.

3 Related works

Referring expression generation (REG) is often framed as an optimization problem where a speaker must identify the lowest-cost (e.g., shortest) utterance which uniquely distinguishes a target referent in the context of distractors. While this optimization is typically modeled at a whole-utterance level, e.g., via the rational speech acts (RSA) model (Frank & Goodman, 2012; Goodman & Frank, 2016), Cohn-Gordon et al. (2018) achieve improved communicative success with LSTM-based speakers using incremental character-level RSA. Modern VLMs are typically able to accurately describe target referents in isolation, but they very often generate pragmatically suboptimal utterances that are unnecessarily verbose (Kapur et al., 2026) or ambiguous with respect to the entire image, resulting in both low communicative success and inefficient interaction (Ma et al., 2025).

Recent work finetunes existing VLMs to produce more pragmatically optimal references by learning from communicative success through interaction with listeners (Willemsen & Skantze, 2024; Tang et al., 2024; Gul & Artzi, 2024). During learning, references are sampled from the speaker policy (i.e., REG), then sent to a listener, who selects the point or region of the image that the sampled reference most likely refers to (i.e., REC). References that result in communicative success are kept as positive examples, while references for which the listener failed to identify the target are kept as negative examples. This prior work varies in how positive and negative examples are converted to learning signals; for example, Gul & Artzi (2024) cast learning as a contextual bandit problem, assigning positive rewards to positive examples and negative rewards to negative ones. However, in all existing work, the learning signal is sparse: each reference is paired with only one piece of feedback, dependent on the final *outcome* of a reference game. Here, we use observations of the *process* of comprehension as a learning signal, as evidenced by the trajectory of listener gaze during reference comprehension.

[3] We say a coordinate $(x, y) \in b = (b_x, b_y, b_w, b_h)$ (a "hit") if $b_x \leq x \leq b_x + b_w$ and $b_y \leq y \leq b_y + b_h$.

In machine learning, human gaze data has been modeled and used as a behavior of interest both as a communication device and learning signal, particularly in natural language processing (NLP) and human-robot interaction (HRI). Gaze and attention data have also been used to guide language generation in descriptive visual tasks like image captioning, video captioning, and video summarization (Yu et al., 2017; Chen & Zhao, 2018; Xu et al., 2015). In HRI, gaze of human operators in jointly embodied environments with robots has been studied as communicative signal for establishing joint attention (Huang & Mutlu, 2016; Rüppel et al., 2025), but embodied gaze has not been explored as a signal from which general language-using agents can learn. In NLP, human gaze patterns during reading of written text have been used as a learning signal to improve performance on text-related tasks (Mishra et al., 2016; Barrett & Hollenstein, 2020; Kiegeland et al., 2024; Papadopoulos, 2025; Bondar et al., 2025; Lopez-Cardona et al., 2025). However, this work is limited to non-embodied settings, where gaze provides a signal more about human processing of the text itself rather than its grounding to the world.

Existing work on gaze estimation focuses mainly on free-viewing tasks (Kümmerer et al., 2022; Kümmerer & Bethge, 2023; Yin et al., 2024) and visual search (Hu et al., 2021b; Cartella et al., 2025a;b; Mondal et al., 2025) conditioned on simple single-word references. Some recent works study gaze estimation during more complex language understanding tasks, such as visual question answering (Chen et al., 2021). Most similar to our approach, Mondal et al. (2024) study estimating human gaze during REC by collecting scanpaths over RefCOCO (Kazemzadeh et al., 2014) and training a corresponding gaze-prediction model with a single-stream VLM architecture.

4 Estimating listener gaze

We train an incremental listener policy, which we refer to as Molmo-REC-Gaze, to estimate human gaze scanpaths during reference comprehension. We initialize the policy's parameters as Molmo-7B-D (Deitke et al., 2024) and finetune it using examples adapted from the RefCOCO-Gaze dataset (Mondal et al., 2024). This approach results in gaze scanpath estimation superior to previous approaches such as ART (Mondal et al., 2024).

4.1 Data

RefCOCO (Kazemzadeh et al., 2014) is a popular corpus used for the study of multimodal reference. Mondal et al. (2024) introduce RefCOCO-Gaze, which labels a subset of examples from RefCOCO with scanpaths of human participants' gaze during reference comprehension. We adapt this dataset to train a neural model that estimates the gaze scanpath of a human listener for arbitrary images and referring expressions. To acquire each scanpath, they display the image (without the bounding box annotation) to the participant, and instruct the participant to focus their attention to the center of the image. The human-written referring expression is then verbalized with word onset timestamps recorded. During reference comprehension, an infrared video-based eye-tracker records participant scanpaths, each of which comprises a sequence of fixations paired with timestamps. Each scanpath in RefCOCO-Gaze demonstrates successful referential comprehension.

Word and fixation alignment To create demonstration data for our incremental listener policy, we must pair each word in the referring expression to a single fixation. Word onset and fixation timestamps included in RefCOCO-Gaze allow us to identify, for each fixation, which word was most recently heard by the participant. We apply a 200ms delay to the timestamp of each word to account for auditory processing time (Kirchner & Thorpe, 2006). There may be multiple fixations per word; when this happens, we select the final fixation before the subsequent word onset. In some cases, there are no corresponding fixations for a word, for example if the word is very short, resulting in a null point $\emptyset$. We convert the RefCOCO-Gaze training set, which contains $D_T = 16{,}982$ scanpaths across 1,799 RefCOCO examples, into a set of demonstrations $\mathcal{D}_{\text{G}}^{\text{train}} = \left\{x^{(i)}, g^{(i)}\right\}_{i=1}^{D_T}$, where $\left|x^{(i)}\right| = \left|g^{(i)}\right|$. For each example, $x_0 = \texttt{BOS}$, g_0 indicates listener gaze before the audio is played, and $x_T = \texttt{EOS}$.

The final fixation $g_T \in b$ is always in the target bounding box. We also convert the validation set into a dataset $\mathcal{D}_{\text{G}}^{\text{val}}$ comprising 869 scanpaths across 92 RefCOCO examples.

4.2 Finetuning for gaze estimation

Most existing VLMs are trained to perform REC by mapping from an image and a referring expression to bounding box coordinates. The Molmo family of VLMs (Deitke et al., 2024) are instead trained to perform REC by mapping form an image and referring expression to a set of point coordinates, which makes it an ideal initialization for estimating gaze. We initialize our gaze listener policy as Molmo-7B-D and optimize it to minimize the token-level loss of gaze sequences included in $\mathcal{D}_{\text{G}}$, i.e., we approximate

$$\phi = \arg\max_{\phi'} \left(-\sum_{i=1}^{D_T} \sum_{t=0}^{|x^{(i)}|} \log \pi_{\phi'}^{\text{G}} \left(g_t^{(i)} \mid x_{\leq t}^{(i)}, g_{<t}^{(i)} \right) \right) .$$

Words in x may be tokenized into multiple subwords; we label each token with the same fixation label. We train for 4 epochs and select the best checkpoint based on loss on the validation set $\mathcal{D}_{\text{G}}^{\text{val}}$. We refer to the final trained policy, which selects $g_t \approx \arg\max \pi_{\phi}^{\text{G}}(\cdot \mid x_{\leq t}, g_{<t})$, as Molmo-REC-Gaze. Figure 5 compares scanpaths from human participants from the validation set $\mathcal{D}_{\text{G}}^{\text{val}}$ with scanpaths estimated by Molmo-REC-Gaze.

Results Mondal et al. (2024) release ART, a VLM trained on RefCOCO-Gaze to estimate human listener gaze, which we evaluate as a baseline to Molmo-REC-Gaze. We report two metrics: (a) path distance, based on dynamic time warping (Senin, 2008), compares a human gaze scanpath with an estimated scanpath; and (b) REC accuracy, the frequency at which an estimated scanpath ends in the target bounding box (REC accuracy). On the validation set $\mathcal{D}_{\text{G}}^{\text{val}}$,[4] Molmo-REC-Gaze significantly outperforms ART, reducing path distance to the closest human scanpath from 55.9 to 39.9 and increasing REC accuracy from 66.7 to 84.4%. See Appendix A for implementation details, formatting changes, and qualitative examples of gaze prediction.

5 Learning from estimated gaze

Our overall learning setup follows existing work on continual learning from interaction (Kojima et al., 2021; Suhr & Artzi, 2023; Gul & Artzi, 2024), where training iterates between deploying the policy being optimized in interaction with another language-using agent (Figure 1a), and optimizing the policy based on observations of the interaction (Figure 1b). In our work, the policy being optimized is a VLM π_{θ}^{S} that maps from images and bounding boxes to a referring expression (i.e., REG), and the listener agent is Molmo-REC-Gaze, which maps from images and referring expressions to an estimated word-aligned scanpath.

Deploying the speaker policy We assume access to a dataset $\mathcal{D}_{\text{I}}$ of images, where each image is annotated with object bounding boxes. In practice, $\mathcal{D}_{\text{I}}$ is the validation split of COCO-2014 (Lin et al., 2015).[5] We randomly split $\mathcal{D}_{\text{I}}$ into training $\mathcal{D}_{\text{I}}^{\text{train}}$ (95%, 38,131 images) and validation $\mathcal{D}_{\text{I}}^{\text{val}}$ (5%, 2,006) splits. During each training interaction, we sample an image $\mathcal{I}$ from $\mathcal{D}_{\text{I}}^{\text{train}}$, then sample a bounding box b whose size is between 1 and 30% of the image's area. This removes target referents that are trivially easy to describe (too large), and ones which are too difficult to perceive (too small). We then sample a referring expression from the current speaker policy; i.e., $x \sim^{\tau} \pi_{\theta}^{\text{S}}(\cdot \mid \mathcal{I}, b)$. Bounding boxes are annotated on the image as a transparent rectangle with a 3-pixel red border. We subsequently estimate a gaze scanpath g from Molmo-REC-Gaze, conditioned on the image $\mathcal{I}$ and sampled referring expression x. To avoid a listener spuriously identifying the target referent, we set the initial pre-reference fixation g_0 to a random coordinate outside of the bounding box b. Each interaction results in an example $(\mathcal{I}, b, x, g)$ such that $|x| = |g|$.

[4] The test set of RefCOCO-Gaze is not publicly available.

[5] We use the validation set to avoid overlap with images on which Molmo-REC-Gaze was trained, as RefCOCO images (thus, RefCOCO-Gaze images) were taken from the COCO-2014 training set.

Optimizing policy from interaction observations We experiment with four reward functions that transform our interaction observation $(\mathcal{I}, b, x, g)$ into a learning signal. Each reward function assigns a reward r_t to each of the generated tokens $x_{t\geq 1}$, taking into account whether fixations in the estimated gaze scanpath are in the bounding box.

Sequence-level last-point hit (SEQLPHIT). Each token x_t receives a positive reward only if the final non-null fixation $g_{T'}$ is in the bounding box, where $T' = \max\{t \leq T : g_t \neq \emptyset\}$. This function rewards all tokens only if the listener's gaze settles on the intended target at the end of the reference.

$$r_t^{\text{SeqLPHit}} = \mathbb{I}[g_{T'} \in b]$$

Sequence-level any hit (SEQANYHIT). Each token x_t receives a positive reward if any fixations in the gaze scanpath are within the bounding box b. We use this reward to identify if the speaker provided sufficient information to ground the referent at any point during the utterance.

$$r_t^{\text{SeqAnyHit}} = \mathbb{I}[\exists k \in \{1, \ldots, T\} : g_k \in b]$$

Before first hit (BEFOREFIRSTHIT). If the interaction demonstrates communicative success, we identify the timestep of the first hit $\tau = \min\{k : g_k \in b\}$; otherwise, we set $\tau = 0$. Then, we assign a positive reward to x_τ and all tokens in its prefix $x_{1\leq t<\tau}$. For successful interactions, we assign a positive reward to the `EOS` token x_T. The goal of this function is to only reinforce the tokens necessary to reach the earliest point of understanding.

$$r_t^{\text{BeforeFirstHit}} = \begin{cases} 1 & \text{if } 1 \leq t \leq \tau \\ 1 & \text{if } t = T \text{ and } \tau > 0 \\ 0 & \text{otherwise} \end{cases}$$

Distance-based shaping (SHAPING). We experiment with reward shaping (Ng et al., 1999) using a potential function based on the distance of each fixation g_t from the bounding box b relative to the initial fixation g_0. We define the distance metric $d(g, b)$ to be the Euclidean distance between the fixation point g and the closest point in the bounding box b. Thus, the reward for each token x_t is the change in relative distance from the fixation associated with the previous token. This rewards tokens that progressively guide the listener's gaze closer to the target, even before a "hit" occurs.

$$r_t^{\text{Shaping}} = \frac{d(g_{t-1}, b) - d(g_t, b)}{d(g_0, b)}$$

After assigning rewards, we use REINFORCE (Sutton et al., 1999) to compute and apply gradients to policy parameters:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{(\mathcal{I},b,x,t)}(r_t - \underbrace{\beta \text{KL}\left(\pi_\theta^{\text{S}}, \mathcal{I}, b, x, t\right)}_{\text{KL penalty}}) \underbrace{\nabla_\theta(\log \pi_\theta^{\text{S}}(x_t \mid \mathcal{I}, b, x_{<t}))}_{\text{Policy gradient}} + \underbrace{\alpha \nabla_\theta H\left(\pi_\theta^{\text{S}}(x_t \mid \mathcal{I}, b, x_{<t})\right)}_{\text{Entropy bonus}},$$

where $\text{KL}\left(\pi_\theta^{\text{S}}, \mathcal{I}, b, x, t\right) = D_{\text{KL}}\left(\pi_\theta^{\text{S}}(x_t \mid \mathcal{I}, b, x_{<t}) \,||\, \pi_{\theta_0}^{\text{S}}(x_t \mid \mathcal{I}, b, x_{<t})\right)$ is the Kullback-Leibler distance between the current policy with parameters θ and the initial policy θ_0.

6 Experimental Setup

Our experiments focus on fine-tuning a Molmo (Deitke et al., 2024) model as a speaker. Evaluation targets two metrics: reference accuracy and efficiency. We compare our approach with learning from interaction with a non-gaze listener model, referred to as Molmo-REC.

Speaker model training We initialize the speaker policy π^{S} with θ_0 as Molmo-7B-D-0924 (Deitke et al., 2024). When sampling, we use temperature $\tau = 0.8$, and randomly select a prompt from the brief prompts in Ma et al. (2025) for each episode. We linearly normalize rewards within-batch to span $[-1, 1]$. We use KL penalty coefficient of $\beta = 0.02$ and an entropy regularization coefficient of $\alpha = 0.01$. In all experiments, we finetune for 1,000 optimizer steps using an effective batch size of 64 and learning rate of 1e-5. We apply LoRA finetuning (Hu et al., 2021a) (rank 8; alpha 16) and checkpoint parameters every 2,000 episodes, selecting the best checkpoint for each run using the reward on the held-out validation set $\mathcal{D}_{\text{I}}^{\text{val}}$.

System variations In addition to the models trained in the main experiments using the four different reward functions (SEQLPHIT, SEQANYHIT, BEFOREFIRSTHIT, and SHAPING;

Speaker	d_{π^s} (Human) ↓	d_{π^s} (Qwen) ↓	Acc. ↑	Length	Time to target (s) ↓
MOLMO	1.41	1.41	75.2	15.4	8.22
GAZE-BEFOREFIRSTHIT	1.11	**0.81**	73.3	4.2	4.86
GAZE-SEQANYHIT	**0.88**	0.95	80.0	9.5	7.04
GAZE-SEQLPHIT	1.07	1.03	77.7	11.0	8.29
GAZE-SHAPING	<u>0.99</u>	0.87	75.5	4.0	5.13
REC-BEFOREFIRSTHIT	1.49	0.99	66.8	2.0	3.98
REC-SEQANYHIT	1.81	1.85	85.0	24.4	11.26
REC-SUCCESS	1.29	1.41	86.0	18.1	8.79
REC-SHAPING	1.08	<u>0.86</u>	73.9	4.4	4.97
HUMAN	0.00	0.00	92.5	3.4	3.87

Table 1: Human evaluation averaged over all datasets with d_{π^s} from Qwen2.5-VL for comparison. Time to target refers to the average time after speech onset when the human evaluators can accurately identify the target by clicking on it.

see Section 5), we compare with (a) the pre-trained Molmo model without any fine-tuning (MOLMO), which serves as a baseline, and (b) human-written references (HUMAN), which serve as an upper bound. We also experiment with two additional interaction settings during learning, where the listener is a pre-trained Molmo-7B-D model. As with our proposed models, we compute a scalar reward for each token x_t based on observations of listener comprehension, and apply REINFORCE. Interaction data and rewards for examples $(\mathcal{I}, b, x)$ are computed as follows:

Communicative success (REC-SUCCESS). We condition a Molmo model on image $\mathcal{I}$ and referring expression x and instruct it to point at the object being referred to, resulting in a single coordinate g_{REC}. This method removes any signals of incremental comprehension.

$$\boxed{r_t^{\text{REC-Success}} = \mathbb{I}[g_{\text{REC}} \in b]}$$

Iterative prompting (REC-ITER). For each prefix in a referring expression $x_{\leq t}$, we condition a Molmo model on image $\mathcal{I}$ and prefix $x_{\leq t}$, and instruct it to point at the object being referred to. This results in a sequence of coordinates $\tilde{g}$. This allows us to use the same reward functions we use with π^G in an effort to disentangle effects of the listener policy and reward design. Policies using this listener are indicated by the prefix REC-.

6.1 Evaluation

Our goal is to train VLMs to generate referring expressions that are both *effective* at communicating an intended referent, and *efficient* in doing so. We report two core metrics: rate of communicative success and reference length. We evaluate across four referring expression datasets, using both human and neural model listeners to measure communicative success.

Metrics For each test dataset $\mathcal{D}_{\text{I}}^{\text{test}} = \left\{\mathcal{I}^{(i)}, b^{(i)}, \hat{x}^{(i)}\right\}_{i=1}^{D_I}$ with D_I examples (each also containing a human-written reference $\hat{x}^{(i)}$), a speaker policy π^{S}, and listener policy π^{REC}, we report the *rate of communicative success* as the proportion of examples in the test set for which the sampled referring expression $x^{(i)}$ results in the listener policy π^{REC} identifying the target reference circumscribed by $b^{(i)}$ in image $\mathcal{I}^{(i)}$. We compute the average *reference length* as the mean number of words. During development of system variants and when choosing hyperparameters, and before deploying models as speakers in interaction with human listeners, we target a metric $d_{\pi^{\text{S}}}$ that combines both core metrics using their distance to the results of human-written references relative to the base model, MOLMO. Metrics for human-written references are computed as:

$$A_{\text{HUMAN}} = \frac{1}{D_I}\sum_{i=1}^{D_I}\mathbb{I}\left[\arg\max \pi^{\text{REC}}\left(\hat{x}^{(i)}, \mathcal{I}^{(i)}\right) \in b^{(i)}\right] \qquad L_{\text{HUMAN}} = \frac{1}{D_I}\sum_{i=1}^{D_I}\left|\hat{x}^{(i)}\right| \ .$$

The normalized distance to the upper-bound performance for some policy π^{S} is then

$$d_{\pi^{\text{S}}} = \sqrt{\left(\max\left(0, \frac{A_{\text{HUMAN}} - A_{\pi^{\text{S}}}}{A_{\text{HUMAN}} - A_{\text{MOLMO}}}\right)\right)^2 + \left(\max\left(0, \frac{L_{\pi^{\text{S}}} - L_{\text{HUMAN}}}{L_{\text{MOLMO}} - L_{\text{HUMAN}}}\right)\right)^2}.$$

When reporting results, we use three random seeds to train three models per system and report the average on all metrics across the three trials.

Datasets We evaluate generated references on four different test sets $\mathcal{D}_{\text{I}}^{\text{test}}$: (a) RefCOCO Test A (1,975 examples), a subset of the RefCOCO test set whose target referents are people; (b) RefCOCO Test B (1,810 examples), a subset of the RefCOCO test set whose target referents are not people; (c) RefOI Co-Occurrence (993 examples) (Ma et al., 2025), a benchmark designed to elicit referring expressions that successfully distinguish between multiple objects resembling the target referent; and (d) RefOI Single Presence (492 examples), which contains simple images that contain a single salient target referent and few to no distractors.

Listener policies We evaluate speaker policy performance in interaction with two listener policies π^{REC}: QWENVL and HUMAN. QWENVL uses the Qwen-VL-2.5-7B model (Bai et al., 2025) to map from an image and referring expression to a bounding box $\hat{b}$.[6] We consider the proposed bounding box $\hat{b}$ to be successful if its intersection over union (IoU) with the target bounding box b is greater than 0.5. HUMAN reflects the performance of speaker models through interaction with human listeners. First, we sample 200 examples from each of the four test sets. In each trial, the image is displayed, and the referring expression is verbalized using Web Speech API's Speech Synthesis [7]. Participants are instructed to click on the region of the image they believe to be described by what they hear. If a participant clicks an image multiple times, we take their final click to be their guess at the target referent. Participants are also given the option to reveal the text of the reference after the audio has played. We collect three responses $\mathcal{P} = \{p_1, p_2, p_3\}$ from three participants per image-reference pair; we consider the set of responses $\mathcal{P} \in b$ if the majority of guesses are in the bounding box; i.e., $\sum_{p \in \mathcal{P}} \mathbb{I}(p \in b) \geq 2$. All three annotators for an example agreed on 77.4% of examples. Statistics on the rate of early clicks and text reveals can be found in Appendix B and the full details and annotator instructions are in Appendix E.

7 Results

Table 1 shows results for all system variations. Models trained with our gaze-based listener Molmo-REC-Gaze significantly reduce the gap $d_{\pi^{\text{S}}}$ from MOLMO to the performance of HUMAN references. Almost all gaze-based models result in references significantly shorter

[6] Although CogVLM-Grounding was found to be the best-performing REC model in Ma et al. (2025), we find it underperforms compared to Qwen-VL (see B.3).

[7] https://developer.mozilla.org/en-US/docs/Web/API/SpeechSynthesis

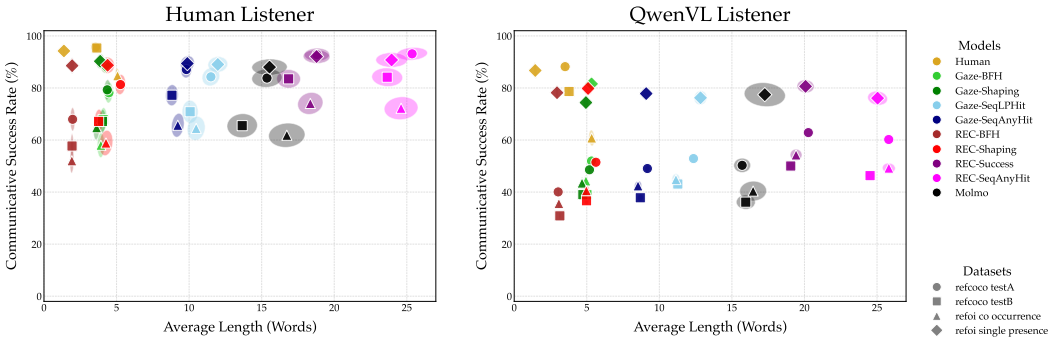


Figure 2: Comparison of communicative success and average reference length in words with human and Qwen2.5-VL listeners with 95% confidence ellipses.

Example 1: RefCOCO Test A



Example 2: RefCOCO Test B



Speaker	Reference: Example 1	Reference: Example 2
Molmo	Man in pink shirt with short brown hair, looking left.	A white bowl with a red border containing a brownish-red sauce, located in the bottom left corner of the image.
REC-SeqAnyHit	A man wearing a pink shirt and gray pants, standing on the far right, with his left hand raised to his chin.	In the bottom left corner, there's a small white bowl containing a red sauce with a white spoon inside. This bowl is highlighted by a red box, making it stand out from the rest of the image.
Gaze-SeqAnyHit	Man wearing pink shirt and gray pants.	Bottom-left corner, white bowl with red border containing brownish-orange sauce.
Gaze-Shaping	Man in pink shirt.	Bottom-left
Human	pink shirt	bowl on bottom left

Figure 3: Examples of generated referring expressions conditioned on an image and bounding box from several different speaker policies. See Appendix D for examples from all speaker models.

and referential success around or higher than MOLMO, demonstrating that learning from listeners that estimate the process of reference comprehension yields improvements in pragmatic competence. When paired with human listeners, these models achieve up to an 4.8% absolute improvement in communicative success (GAZE-SEQANYHIT) and a reduction from 15.4 down to 4.0 words per reference (GAZE-SHAPING). On the other hand, while speakers trained with non-gaze-based listeners achieve even higher boosts in communicative success up to a 10.8% absolutely improvement (REC-SUCCESS), this comes at a serious cost to reference efficiency, up to references with an average of 24.4 words (REC-SEQANYHIT). This is also reflected in absolute time for human listeners to identify target referents; while REC-SEQANYHIT outperforms GAZE-SEQANYHIT in communicative success by 5%, identifying target referents takes more than 4 seconds on average longer than GAZE-SEQANYHIT. Although REC-SHAPING performs competitively with the machine listener, the communicative success suffers with human listeners, indicating the presence of some behavior that is less intelligible to human listeners. Further discussion of such behaviors can be found in Appendix D. Our results illustrate a tradeoff between efficiency and efficacy in reference generation (Figure 7). Overall, we find that training with listeners that estimate the process of reference comprehension, rather than only the outcome, results in speakers that generate much more humanlike references.[8]

Beyond main results, we analyze the shifts in referential behavior induced by gaze-based and REC-based learning. We consider the referential strategy of each speaker model using a list of keywords to tag each generated reference as utilizing spatial reference, attribute-based reference, or both (Figure 4). We find that the GAZE-SHAPING policy is the only machine speaker to use mostly spatial reference; its distribution of referential strategies most closely resembles the human distribution. While the accuracy and length of references from GAZE-SHAPING and REC-SHAPING are similar for machine listeners, human evaluation and the differences in referential strategy demonstrate that gaze-based rewards result in more humanlike references.

We also conduct an analysis of the patterns observed in Ma et al. (2025), noting what portions of references are redundant, or if the reference is ambiguous or misleading for a subset of 50 examples (Appendix D.2). Our analysis reveals that gaze-based optimization reduces redundancy, but also produces more ambiguous expressions. While the MOLMO baseline and REC-SUCCESS policies generate an average of 13.35 and 15.01 unnecessary words per

[8]Full results per dataset and all model-based evaluation can be found in Appendix B

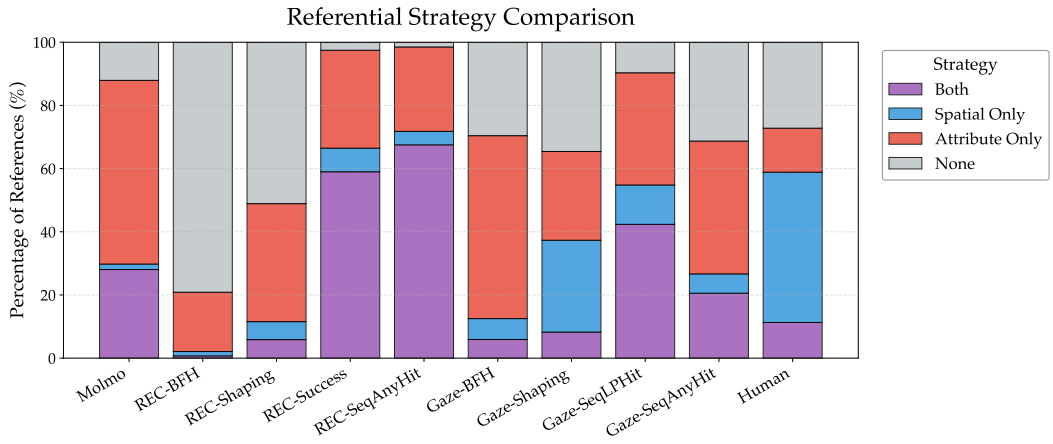


Figure 4: Comparison of referential strategies utilized by different speakers. The keywords used for this analysis can be found in Appendix D.1.

reference, respectively, GAZE-SEQANYHIT nearly halves this redundancy to 7.85 words, and GAZE-SHAPING further minimizes it to just 3.26 words.

8 Conclusion

In this work, we demonstrate that VLMs can be finetuned to generate more pragmatically optimal referring expressions using gaze as an implicit signal of listener comprehension. Policies that learn from gaze reduce the gap between base VLM outputs and human-like performance, reducing average length by as much as 11.4 words while maintaining or improving communicative success rates. Furthermore, the GAZE-SHAPING policy is the only speaker policy to prioritize spatial reference, mirroring human behavior without explicit feedback. These results demonstrate that learning from signals provided from a human-like listener result in more human-like generations. Nonetheless, a tradeoff remains; speaker policies trained with REC listeners achieve the highest accuracy rates, but do so by producing verbose references. Overall, our findings establish that gaze-based insights into the process of comprehension offer a valuable implicit signal for advancing pragmatic language generation.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research was supported by an Amazon Research Award, a gift from Google, a Technical AI Safety Research award from Coefficient Giving, and an NVIDIA Academic Grant Program award.

Ethics Statement

This work focuses exclusively on English, potentially reinforcing linguistic patterns and cultural norms from high-resource languages; our approach may inadvertently marginalize low-resource or non-standard dialects. We acknowledge that generating human-like expressions requires ongoing oversight to prevent the reinforcement of systemic linguistic biases.

All human evaluation followed ethical research practices, with 586 participants completing the study after quality filtering. Participants were compensated at an average rate of $19.67 USD per hour. Interaction data, including clicks and timestamps, was fully anonymized to protect participant privacy.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Maria Barrett and Nora Hollenstein. Sequence labelling and sequence classification with gaze: Novel uses of eye-tracking data for natural language processing. *Language and Linguistics Compass*, 14(11):1–16, 2020.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. Paligemma: A versatile 3b vlm for transfer, 2024. URL https://arxiv.org/abs/2407.07726.

Anna Bondar, David Robert Reich, and Lena Ann Jäger. AlEYEgnment: Leveraging Eye-Tracking-While-Reading to align language models with human preferences. In Cengiz Acarturk, Jamal Nasir, Burcu Can, and Cagrı Coltekin (eds.), *Proceedings of the First International Workshop on Gaze Data and Natural Language Processing*, pp. 58–70, Varna, Bulgaria, September 2025. INCOMA Ltd., Shoumen, BULGARIA. URL https://aclanthology.org/2025.gaze4nlp-1.8/.

Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16206–16216, October 2025a.

Giuseppe Cartella, Vittorio Cuculo, Alessandro D'Amelio, Marcella Cornia, Giuseppe Boccignone, and Rita Cucchiara. Modeling human gaze behavior with diffusion models for unified scanpath prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16206–16216, 2025b.

Shi Chen and Qi Zhao. Boosted attention: Leveraging human attention for image captioning. In *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pp. 72–88. Springer International Publishing, Cham, 2018.

Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10876–10885, 2021.

Herbert H Clark and Meredyth A Krych. Speaking while monitoring addressees for understanding. *Journal of memory and language*, 50(1):62–81, 2004.

Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. *Cognition*, 22(1):1–39, 1986.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. Pragmatically informative image captioning with character-level inference. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 439–443, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2070. URL https://aclanthology.org/N18-2070/.

Robert Dale and Ehud Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Michael C Frank and Noah D Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998, 2012.

Noah D Goodman and Michael C Frank. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829, 2016.

Herbert P Grice. Logic and conversation. In *Speech acts*, pp. 41–58. Brill, 1975.

Mustafa Omer Gul and Yoav Artzi. Cogen: Learning from feedback with coupled comprehension and generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 12966–12982, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021a. URL https://arxiv.org/abs/2106.09685.

Zhiming Hu, Andreas Bulling, Sheng Li, and Guoping Wang. FixationNet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Trans. Vis. Comput. Graph.*, 27(5): 2681–2690, May 2021b.

Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pp. 83–90. IEEE, 2016.

Julian Jara-Ettinger and Paula Rubio-Fernandez. The social basis of referential communication: Speakers construct physical reference based on listeners' expected visual search. *Psychological review*, 129(6):1394, 2022.

Rhea Kapur, Robert Hawkins, and Elisa Kreiss. When more words say less: Decoupling length and specificity in image description evaluation, 2026. URL https://arxiv.org/abs/2601.04609.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/.

Samuel Kiegeland, D R Reich, Ryan Cotterell, Lena A J¨ager, and E Wilcox. The pupil becomes the master: Eye-tracking feedback for tuning LLMs. *LLMs and Cognition at ICML*, 2024.

Holle Kirchner and Simon J Thorpe. Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision research*, 46(11):1762–1776, 2006.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. Continual learning for grounded instruction generation by observing human following behavior. *Transactions of the Association for Computational Linguistics*, 9:1303–1319, 2021. doi: 10.1162/tacl_a_00428. URL https://aclanthology.org/2021.tacl-1.77/.

Matthias Kümmerer and Matthias Bethge. Predicting visual fixations. *Annu. Rev. Vis. Sci.*, 9 (Volume 9, 2023):269–291, September 2023.

Matthias Kümmerer, Matthias Bethge, and Thomas S A Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *J. Vis.*, 22(5):7, April 2022.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Angela Lopez-Cardona, Carlos Segura, Alexandros Karatzoglou, Sergi Abadal, and Ioannis Arapakis. Seeing eye to ai: Human alignment via gaze-based response rewards for large language models, 2025. URL https://arxiv.org/abs/2410.01532.

Ziqiao Ma, Jing Ding, Xuejun Zhang, Dezhi Luo, Jiahe Ding, Sihan Xu, Yuchen Huang, Run Peng, and Joyce Chai. Vision-language models are not pragmatically competent in referring expression generation. *arXiv [cs.CL]*, July 2025.

Abhijit Mishra, Diptesh Kanojia, and Pushpak Bhattacharyya. Predicting readers' sarcasm understandability by modeling gaze behavior. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Mar. 2016. doi: 10.1609/aaai.v30i1.9884. URL https://ojs.aaai.org/index.php/AAAI/article/view/9884.

Sounak Mondal, Seoyoung Ahn, Zhibo Yang, Niranjan Balasubramanian, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. Look hear: Gaze prediction for speech-directed human attention. *arXiv [cs.CV]*, July 2024.

Sounak Mondal, Naveen Sendhilnathan, Ting Zhang, Yue Liu, Michael Proulx, Michael Louis Iuzzolino, Chuan Qin, and Tanya R Jonker. Gaze-language alignment for zero-shot prediction of visual search targets from human gaze scanpaths. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2738–2749, 2025.

Andrew Y. Ng, Daishi Harada, and Stuart J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 278–287, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1558606122.

Nikolas Papadopoulos. Eye-tracking as implicit feedback for aligning large language models and enhancing human-ai teaming. In *Proceedings of the 2025 Symposium on Eye Tracking Research and Applications*, ETRA '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714870. doi: 10.1145/3715669.3727334. URL https://doi.org/10.1145/3715669.3727334.

Jens V Rüppel, Andrey Rudenko, Tim Schreiter, Martin Magnusson, and Achim J Lilienthal. Gaze-supported large language model framework for bi-directional human-robot interaction. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 677–684. IEEE, 2025.

Pavel Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, 855(1-23):40, 2008.

Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'99, pp. 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

Zineng Tang, Lingjun Mao, and Alane Suhr. Grounding language in multi-perspective referential communication. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19727–19741, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1100. URL https://aclanthology.org/2024.emnlp-main.1100/.

Bram Willemsen and Gabriel Skantze. Referring expression generation in visually grounded dialogue with discourse-aware comprehension guiding. In Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito (eds.), *Proceedings of the 17th International Natural Language Generation Conference*, pp. 453–469, Tokyo, Japan, September 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.inlg-main.38. URL https://aclanthology.org/2024.inlg-main.38/.

Jia Xu, Lopamudra Mukherjee, Yin Li, Jamieson Warner, James M Rehg, and Vikas Singh. Gaze-enabled egocentric video summarization via constrained submodular maximization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2235–2244, 2015.

Pengwei Yin, Jingjing Wang, Guanzhong Zeng, Di Xie, and Jiang Zhu. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, pp. 1–17. Springer, 2024.

Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017.

A Gaze prediction details and evaluation

A.1 Point format augmentation

For the Molmo-based architectures, we augment the standard point format to exclude alternate text and quotation marks to ensure simplicity during reward calculation. The Molmo point formatting

$$< \text{point x=}''50.0'' \text{ y=}''50.0'' \text{ alt=}''\text{cat}'' > \text{cat} < /\text{point} >$$

is simplified during finetuning the gaze predictor to:

$$< \text{x=50.0 y=50.0} >$$

This formatting change also applies to other finetuned listener models trained for ablations in following appendices.

A.2 Example of gaze prediction



Gold (Human)	BOS bus on right in back EOS
Molmo-REC-Gaze	BOS bus on right in back EOS

Figure 5: Example of the iterative gaze prediction for gold and finetuned Molmo. The orange sequence represents the human scanpath and the blue is that produced by our gaze predictor model. Each point indicates the location of the gaze and the word heard at the time the fixation was recorded.

A.3 Full gaze prediction results

Below are the full evaluation results for our gaze predictor and ART (Mondal et al., 2024). We include the distance to the gold point both averaged over all gold scanpaths and using the closest gold scanpath at each word index and total. In both cases, the VLM-based predictor is closer to the human scanpath than ART.

Word Index	Molmo Avg Dist	ART Avg Dist
0	2.04	2.01
1	10.56	6.59
2	8.63	6.54
3	9.92	8.56
4	9.39	7.70
5	11.78	9.99
6	2.70	7.23
7	7.34	7.87
8	3.31	11.22
Overall	**43.21**	45.75

Table 3: Average Euclidean distance per word index and total sum of distances, using the closest gold path (Best Match).

Word Index	Molmo Avg Dist	ART Avg Dist
0	4.25	4.43
1	12.17	9.81
2	12.85	12.56
3	13.59	14.08
4	12.10	11.56
5	12.77	13.11
6	5.28	12.83
7	13.12	8.83
8	3.21	10.34
Overall	**62.24**	72.86

Table 2: Average Euclidean distance per word index and total sum of distances, averaged over all gold sequences.

A.4 Predicting gaze with human and machine-generated references

Since the gaze prediction model is trained only on human-generated references, we analyze differences in gaze prediction accuracy between the gold references and machine-generated references resembling those the model would see at the start of training. To do so, we sample 100 examples from each of the four evaluation datasets and generate scanpaths for these 400 examples using the gold references for each as well as references generated by Molmo-7B with no additional training.

Across human and machine-generated references, the gaze predictor achieves higher accuracy and fewer null points on shorter references, although it is difficult to disentangle the length or speaker from the difficulty of the produced references. The difference in performance definitely influences the resulting speaker policies trained with the gaze listener. However, it is intentional that the listener used during speaker training reflects a human listener's reference resolution process, in order to result in a speaker that generates references that are easy for a human listener to understand. The Molmo REC listener excels at reference resolution, in that it is trained to understand references that humans cannot (i.e., coordinates). Our results show that a listener trained on the kinds of expressions people use, and the processes involved in resolving such references, results in the production of more humanlike references.

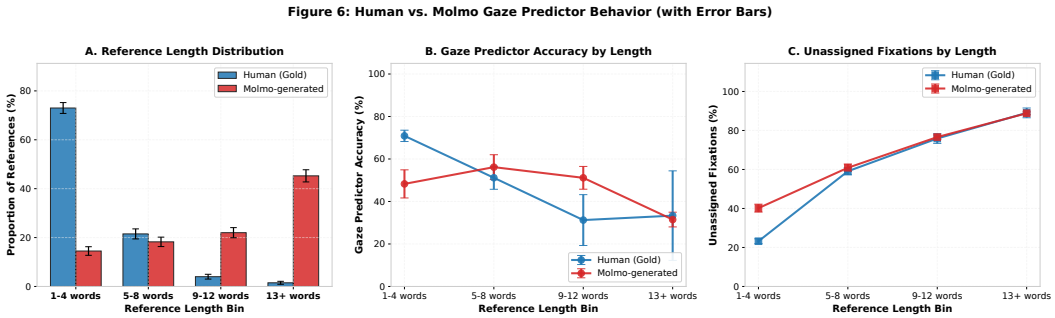


Figure 6: (A) Comparison of length distributions of human and Molmo-generated references, (B) accuracy of the gaze predictor on human-generated (blue) and Molmo-generated references (red) over different reference lengths, and (C) comparison of the proportion of words not assigned fixations.

B Full REG results and evaluator choice

B.1 Full REG results

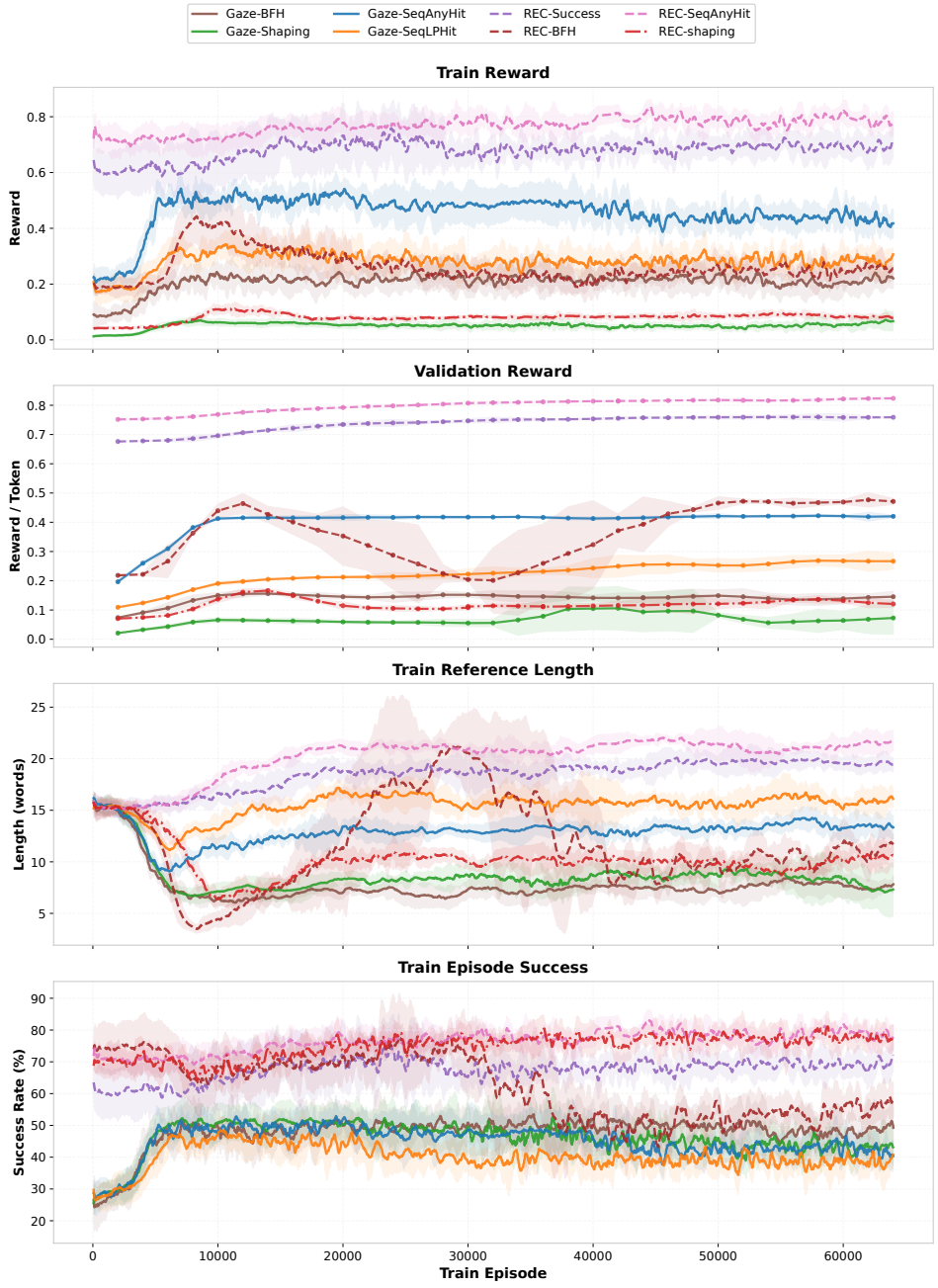
In this section, we report the complete dataset-level evaluation for all speaker variants. Table 4 details the REF performance metrics: accuracy, average utterance length, average time for a human listener to resolve the expression, and d_{π^s}. Additionally, Table 5 presents behavioral metrics— resolution time, reveal rate, and early click rate. Finally, we present the averaged and per-dataset results using the QwenVL listener in Tables 6 and 7, respectively.

Speaker	Acc. ↑	Length	Time (s) ↓	d_{π^s} ↓
RefCOCO testA				
HUMAN (GOLD)	95.67	3.59	3.97	0.00
MOLMO	84.17	15.15	7.99	1.41
REC-BEFOREFIRSTHIT	68.50	2.01	4.23	2.37
REC-SEQANYHIT	93.17	25.27	11.40	1.89
REC-SUCCESS	93.50	18.58	9.21	1.31
REC-SHAPING	81.17	5.27	5.30	1.27
GAZE-SEQANYHIT	87.83	9.87	6.35	0.87
GAZE-SEQLPHIT	85.00	11.53	8.94	1.15
GAZE-SHAPING	79.83	4.36	4.52	1.38
GAZE-BEFOREFIRSTHIT	79.00	4.48	5.12	1.45
RefCOCO testB				
HUMAN (GOLD)	96.17	3.73	4.45	0.00
MOLMO	65.50	13.84	8.19	1.41
REC-BEFOREFIRSTHIT	58.17	1.93	3.99	1.25
REC-SEQANYHIT	83.83	23.56	11.98	2.00
REC-SUCCESS	83.67	16.90	8.76	1.37
REC-SHAPING	67.00	3.78	5.20	0.95
GAZE-SEQANYHIT	77.33	8.84	7.13	0.80
GAZE-SEQLPHIT	71.33	10.08	8.04	1.02
GAZE-SHAPING	67.33	4.06	5.44	0.94
GAZE-BEFOREFIRSTHIT	66.83	4.04	5.17	0.96
RefOI Co-occurrence				
HUMAN (GOLD)	84.83	5.02	4.07	0.00
MOLMO	62.83	17.00	9.45	1.41
REC-BEFOREFIRSTHIT	52.00	1.98	4.08	1.51
REC-SEQANYHIT	72.50	24.70	11.50	1.74
REC-SUCCESS	74.67	18.36	8.66	1.20
REC-SHAPING	59.00	4.21	5.51	1.18
GAZE-SEQANYHIT	65.83	9.24	7.24	0.93
GAZE-SEQLPHIT	65.17	10.48	7.78	1.00
GAZE-SHAPING	64.33	3.63	5.35	0.94
GAZE-BEFOREFIRSTHIT	58.33	3.86	5.28	1.21
RefOI Single Presence				
HUMAN (GOLD)	93.50	1.39	3.28	0.00
MOLMO	88.33	15.57	7.57	1.41
REC-BEFOREFIRSTHIT	88.33	1.98	3.73	1.00
REC-SEQANYHIT	90.67	24.00	10.26	1.69
REC-SUCCESS	92.00	18.68	8.51	1.25
REC-SHAPING	88.33	4.45	4.14	1.02
GAZE-SEQANYHIT	89.00	9.94	6.67	1.06
GAZE-SEQLPHIT	89.33	12.07	8.24	1.10
GAZE-SHAPING	90.33	3.89	4.44	0.64
GAZE-BEFOREFIRSTHIT	88.83	4.29	4.12	0.93

Table 4: Full human evaluation results by dataset.

B.2 Training curves

Fig B.2 shows the train and validation reward, success rate of the reference (i.e., whether any point in the set of produced point was within the target region), and generated reference length in tokens. The shaded area represents the range of values over three random seeds. The rewards and success rates for gaze-based rewards tend to be lower than their REC counterparts. The REC-BEFOREFIRSTHIT training is noticeably the most unstable relative to the other reward configurations.



Speaker	Time (s) ↓	Reveal Rate ↓	Early Click Rate ↓
Overall			
HUMAN (GOLD)	3.87	0.03	0.05
MOLMO	8.22	0.05	0.27
REC-BEFOREFIRSTHIT	3.98	0.04	0.04
REC-SEQANYHIT	11.26	0.05	0.45
REC-SUCCESS	8.79	0.03	0.41
REC-SHAPING	5.04	0.09	0.13
GAZE-SEQANYHIT	7.04	0.03	0.21
GAZE-SEQLPHIT	8.29	0.04	0.32
GAZE-SHAPING	5.13	0.04	0.06
GAZE-BEFOREFIRSTHIT	4.86	0.04	0.06
RefCOCO testA			
HUMAN (GOLD)	3.97	0.03	0.04
MOLMO	7.99	0.05	0.24
REC-BEFOREFIRSTHIT	4.23	0.06	0.02
REC-SEQANYHIT	11.40	0.05	0.36
REC-SUCCESS	9.21	0.04	0.38
REC-SHAPING	5.30	0.07	0.13
GAZE-SEQANYHIT	6.35	0.03	0.27
GAZE-SEQLPHIT	8.94	0.05	0.35
GAZE-SHAPING	4.52	0.05	0.05
GAZE-BEFOREFIRSTHIT	5.12	0.04	0.06
RefCOCO testB			
HUMAN (GOLD)	4.45	0.05	0.04
MOLMO	8.19	0.05	0.17
REC-BEFOREFIRSTHIT	3.99	0.05	0.04
REC-SEQANYHIT	11.98	0.06	0.39
REC-SUCCESS	8.76	0.05	0.31
REC-SHAPING	5.20	0.10	0.10
GAZE-SEQANYHIT	7.13	0.05	0.14
GAZE-SEQLPHIT	8.04	0.05	0.28
GAZE-SHAPING	5.44	0.06	0.06
GAZE-BEFOREFIRSTHIT	5.17	0.05	0.04
RefOI Co-occurrence			
HUMAN (GOLD)	4.07	0.03	0.07
MOLMO	9.45	0.04	0.29
REC-BEFOREFIRSTHIT	4.08	0.03	0.06
REC-SEQANYHIT	11.50	0.06	0.42
REC-SUCCESS	8.66	0.03	0.42
REC-SHAPING	5.51	0.11	0.11
GAZE-SEQANYHIT	7.24	0.04	0.15
GAZE-SEQLPHIT	7.78	0.06	0.26
GAZE-SHAPING	5.35	0.04	0.03
GAZE-BEFOREFIRSTHIT	5.28	0.04	0.05
RefOI Single Presence			
HUMAN (GOLD)	3.28	0.03	0.05
MOLMO	7.57	0.05	0.39
REC-BEFOREFIRSTHIT	3.73	0.03	0.06
REC-SEQANYHIT	10.26	0.05	0.63
REC-SUCCESS	8.51	0.02	0.53
REC-SHAPING	4.14	0.07	0.17
GAZE-SEQANYHIT	6.67	0.03	0.28
GAZE-SEQLPHIT	8.24	0.03	0.39
GAZE-SHAPING	4.44	0.03	0.10
GAZE-BEFOREFIRSTHIT	4.12	0.02	0.11

Table 5: Behavioral efficiency metrics grouped by dataset, including overall averages.

Speaker	Acc. ↑	Length	d_{π^s} ↓
RefCOCO testA			
HUMAN (GOLD)	94.60	3.51	0.00
MOLMO	48.20	15.71	1.41
REC-BEFOREFIRSTHIT	40.07	3.03	1.18
REC-SEQANYHIT	60.20	25.81	1.97
REC-SUCCESS	62.85	20.27	1.53
REC-SHAPING	51.46	5.63	0.95
GAZE-SEQANYHIT	49.05	9.18	1.09
GAZE-SEQLPHIT	52.88	12.35	1.16
GAZE-SHAPING	48.57	5.19	1.00
GAZE-BEFOREFIRSTHIT	51.93	5.33	0.93
RefCOCO testB			
HUMAN (GOLD)	88.00	3.78	0.00
MOLMO	33.50	15.96	1.41
REC-BEFOREFIRSTHIT	30.87	3.14	1.05
REC-SEQANYHIT	46.32	24.52	1.87
REC-SUCCESS	50.02	19.06	1.44
REC-SHAPING	36.67	4.98	0.95
GAZE-SEQANYHIT	37.82	8.69	1.01
GAZE-SEQLPHIT	43.09	11.27	1.03
GAZE-SHAPING	39.01	4.73	0.90
GAZE-BEFOREFIRSTHIT	38.99	5.01	0.90
RefOI Co-occurrence			
HUMAN (GOLD)	92.90	5.34	0.00
MOLMO	33.20	16.46	1.41
REC-BEFOREFIRSTHIT	35.62	3.08	0.98
REC-SEQANYHIT	49.22	25.81	1.98
REC-SUCCESS	54.28	19.42	1.42
REC-SHAPING	40.55	4.97	0.88
GAZE-SEQANYHIT	42.40	8.53	0.89
GAZE-SEQLPHIT	44.81	11.14	0.96
GAZE-SHAPING	43.47	4.67	0.83
GAZE-BEFOREFIRSTHIT	44.31	4.94	0.81
RefOI Single Presence			
HUMAN (GOLD)	80.40	1.45	0.00
MOLMO	67.70	17.27	1.41
REC-BEFOREFIRSTHIT	78.18	2.95	0.20
REC-SEQANYHIT	76.08	25.04	1.53
REC-SUCCESS	80.62	20.07	1.18
REC-SHAPING	79.81	5.09	0.23
GAZE-SEQANYHIT	77.85	9.09	0.52
GAZE-SEQLPHIT	76.22	12.84	0.79
GAZE-SHAPING	74.39	4.94	0.52
GAZE-BEFOREFIRSTHIT	81.57	5.33	0.26

Table 6: QwenVL-based evaluation results by dataset.

Speaker	Acc. ↑	Length	d_{π^s} ↓
Overall (Averaged)			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
HUMAN (GOLD)	88.98	3.52	0.00

Table 7: QwenVL-based evaluation results averaged across all datasets.

B.3 Comparison of machine listener models

We choose to use QwenVL for our main evaluation despite Ma et al. (2025) finding CogVLM-Grounding to be the best model for REC. As shown in 7, the reference resolution rate of CogVLM-Grounding becomes extremely low as references get longer, whereas QwenVL maintains higher success rates across reference lengths. We hypothesize that this difference reflects a training bias in CogVLM-Grounding toward concise, human-written references.

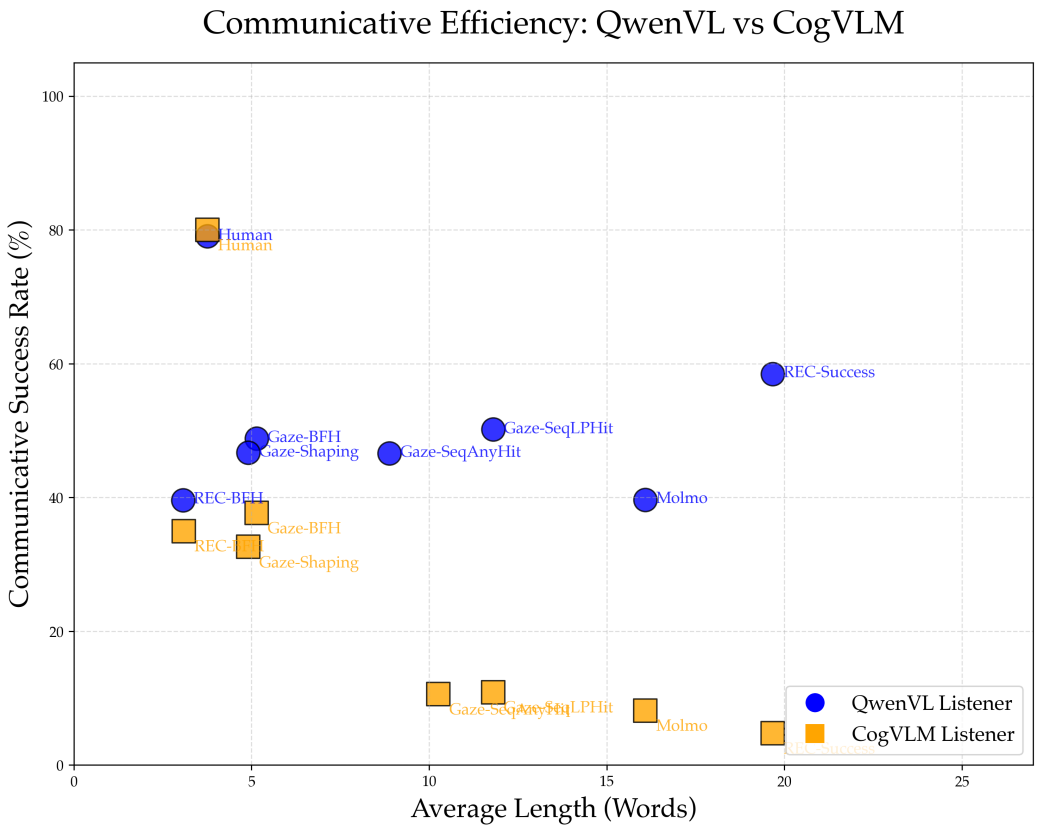


Figure 7: Comparison of REC results with QwenVL and CogVLM-Grounding.

C Speaker model ablations, baselines, and comparisons

C.1 KL penalty ablation

We demonstrate the utility of the KL penalty on the highest performing speaker model with the QwenVL listener (GAZE-BFH) by ablating it. Although training without the KL penalty yields higher validation rewards late in training, it causes entropy collapse and degrades overall evaluation performance (Figure 8). With reference length collapsing to just 1–2 tokens on average, these results confirm that the KL penalty is necessary to mitigate reward hacking. As a result, we apply the KL penalty to the training of all speaker models.

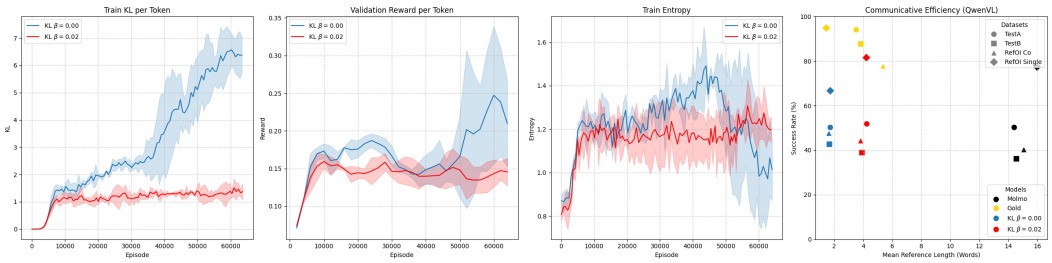


Figure 8: Comparison of GAZE-BFH training performance with and without the KL penalty. Red shows training with a KL penalty ($\beta = 0.02$) and blue shows the ablated training run. For comparison, we include Molmo and human speakers in the rightmost chart.

C.2 REC-Success with length penalty

While the goal of this work is not to demonstrate that learning from gaze results in the most efficient and successful references, we compare to REC-SUCCESS incorporating a length penalty to control for verbosity. Because policy optimization can favor longer sequences to maximize reward, penalizing length allows for a fair comparison at closer sequence lengths. We implement the length penalty as a negative reward, $\lambda = 0.05$. The penalty is applied after batch normalization to each generated token with the exception of the end-of-sequence token. The training objective becomes:

$$\nabla_\theta \mathcal{J}(\theta) = \mathbb{E}_{(\mathcal{I},b,x,t)}\Big(r_t - \underbrace{\beta \mathrm{KL}\left(\pi_\theta^{\mathrm{S}}, \mathcal{I}, b, x, t\right)}_{\text{KL penalty}} - \underbrace{\lambda \mathbb{I}(x_t \neq \mathrm{EOS})}_{\text{Length penalty}}\Big) \times \underbrace{\nabla_\theta \left(\log \pi_\theta^{\mathrm{S}}\left(x_t \mid \mathcal{I}, b, x_{<t}\right)\right)}_{\text{Policy gradient}} + \underbrace{\alpha \nabla_\theta H\left(\pi_\theta^{\mathrm{S}}(x_t \mid \mathcal{I}, b, x_{<t})\right)}_{\text{Entropy bonus}}, \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Speaker	Acc. ↑	Length	d_{π^s} ↓
Overall (Averaged)			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
HUMAN (GOLD)	88.97	3.52	0.00
Length Penalty ($\lambda = 0.05$)			
REC-SUCCESS (LP $\lambda = 0.05$)	61.08	9.50	0.83

Table 8: QwenVL-based evaluation comparing standard models against the length-penalized REC-SUCCESS variant.

The speaker trained with REC-SUCCES and a length penalty outperforms almost all other configurations, but falls just short of GAZE-BEFOREFIRSTHIT. Explicit length penalization successfully reduces verbosity, but gaze-based feedback can also encourage concise, pragmatically optimal expressions without relying on manually tuned penalty weights.

C.3 Listener variant vaselines

C.3.1 Final gaze point listener

To ensure the brevity of gaze-trained speakers is driven by the gaze signal and reward design rather than training data bias, we evaluate a speaker trained with a variant of π^G that predicts only the final fixation point. We then integrate this model with the REC-SUCCESS training configuration. The resulting speaker model results in length and accuracy similar to that of REC-SUCCESS and even slightly longer, demonstrating that the gaze listener's training distribution is not the cause of the shorter references.

Speaker	Acc. ↑	Length	d_{π^s} ↓
Overall (Averaged)			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
HUMAN (GOLD)	88.98	3.52	0.00
Baseline: Final Gaze Point Listener			
GAZE-SUCCESS (FINAL GAZE POINT LISTENER)	61.17	21.47	1.56

Table 9: Evaluation comparing speakers trained on full incremental gaze versus models predicting only the final gaze point.

C.3.2 RefCOCO finetuned listener

To ensure the gaze listener does not outperform the VLM listener solely through indirect exposure to RefCOCO grounding data, we finetune Molmo-7B on the RefCOCO subset

within the RefCOCO-Gaze training set for a fair comparison. Then, similarly to the prior baselines, we train a speaker against this listener with REC-SUCCESS rewards. Once again, we find that exposure to RefCOCO does not result in higher accuracy or shorter references.

Speaker	Acc. ↑	Length	d_{π^s} ↓
Overall (Averaged)			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
HUMAN (GOLD)	88.98	3.52	0.00
Baseline: RefCOCO-Finetuned Listener			
REC-SUCCESS (REFCOCO LISTENER)	59.93	21.64	1.57

Table 10: Performance comparison when training the speaker against a listener finetuned directly on RefCOCO.

C.4 Comparison to SFT

To establish an upper bound, we finetune Molmo-7B-D directly on the RefCOCO validation set (8.8k examples). We consider 1%, 10%, and 100% of this train set. Each model is trained for 4 epochs at learning rate 5e-5. As expected, direct supervision on human-written references yields strong performance. We additionally try training from the RefCOCO 1% checkpoint with the GAZE-SHAPING and GAZE-SEQANYHIT reward setups; however, this degrades performance from the SFT checkpoint.

Speaker	Acc. ↑	Length	d_{π^s} ↓
Overall (Averaged)			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
HUMAN (GOLD)	88.98	3.52	0.00
Comparison to SFT Baselines			
REFCOCO SFT 100%	64.44	3.34	0.59
REFCOCO SFT 10%	64.56	3.17	0.59
REFCOCO SFT 1%	60.27	3.28	0.71
REFCOCO SFT 1% + SHAPING GAZE	55.21	3.79	0.84
REFCOCO SFT 1% + BINARY GAZE	59.54	7.34	0.78

Table 11: Comparison of Supervised Finetuning (SFT) baselines and SFT+RL hybrid models against the default RL/Gaze speakers.

C.5 Generalization to other VLMs

We recreate a subset of the main results, instead training PaliGemma-3B-pt-448 (Beyer et al., 2024) and LLaVA-1.5-7B (Liu et al., 2024) as speaker models. This subset includes GAZE-SHAPING, GAZE-BEFOREFIRSTHIT, GAZE-SEQANYHIT, REC-SHAPING, REC-BEFOREFIRSTHIT, REC-SEQANYHIT, and REC-SUCCESS.

Table 12 reports results; we observe that gains do not cleanly generalize across VLMs. Because hyperparameters were held constant from the Molmo-7B experiments, we attribute this to a lack of model-specific tuning, especially for the KL penalty coefficient (β), and differences in base models and their initial policy entropy.

Speaker	**Acc.** ↑	**Length**	d_{π^s} ↓
HUMAN (GOLD)	88.98	3.52	0.00
Molmo-7B			
MOLMO	45.65	16.35	1.41
REC-BEFOREFIRSTHIT	46.18	3.05	0.99
REC-SEQANYHIT	57.96	25.29	1.84
REC-SUCCESS	61.94	19.70	1.41
REC-SHAPING	52.12	5.17	0.86
GAZE-SEQANYHIT	51.78	8.87	0.95
GAZE-SEQLPHIT	54.25	11.90	1.03
GAZE-SHAPING	51.36	4.88	0.87
GAZE-BEFOREFIRSTHIT	54.20	5.15	0.81
LLaVA-1.5-7B			
LLAVA-1.5	28.32	6.16	1.59
REC-BEFOREFIRSTHIT	37.08	1.92	1.34
REC-SHAPING	44.40	2.04	1.14
REC-SUCCESS	53.03	12.06	1.12
GAZE-BEFOREFIRSTHIT	37.08	1.92	1.31
GAZE-SHAPING	20.51	3.19	1.79
GAZE-SEQANYHIT	39.48	3.02	1.27
PaliGemma-3B			
PALIGEMMA	31.80	3.08	1.48
REC-BEFOREFIRSTHIT	41.29	0.96	1.24
REC-SHAPING	41.08	0.96	1.24
REC-SEQANYHIT	34.17	26.09	2.28
REC-SUCCESS	42.92	1.55	1.19
GAZE-BEFOREFIRSTHIT	32.84	1.05	1.46
GAZE-SHAPING	29.38	0.92	1.56
GAZE-SEQANYHIT	35.11	8.00	1.43

Table 12: Cross-architecture evaluation using LLaVA-1.5-7B and PaliGemma-3B as speaker backbones.

D Quality and qualitative analyses

D.1 Strategy analysis

To consider not only the similarity of length and accuracy to human-generated references, but also stylistic choices, we perform a simple analysis of referring expression strategy. Referring expressions are often separated into spatial or attribute based references; using keyword lists, we show the proportion of references that use spatial or attribute based keywords, both, or neither. The keywords used for each category were:

- **Spatial Keywords:** left, right, top, bottom, middle, center, above, below, next, near, far, between, behind, front, centered, side, corner, upper, lower, background, foreground, back, overhead, touching, overlapping, contacting, abutting, along-

side, bordering, nearest, farthest, furthest, closest, distant, adjacent, beside, inside, outside, within, interior, exterior, embedded, surrounded, among, amongst, amid, under, underneath, beneath, atop, upon, over, edge, margin, boundary, perimeter, diagonal, aligned, stacked, facing, opposite, rear, first, second, third, last, outer, inner, innermost, outermost

- **Attribute Keywords:** red, blue, green, yellow, orange, purple, pink, brown, black, white, gray, grey, silver, gold, tan, beige, maroon, navy, teal, cyan, magenta, violet, dark, light, bright, pale, deep, vivid, transparent, translucent, opaque, multicolored, colorful, neon, turquoise, indigo, cream, charcoal, blonde, brunette, auburn, large, small, big, little, tall, short, long, round, circular, square, rectangular, triangular, oval, spherical, cylindrical, flat, curved, pointed, straight, crooked, hollow, solid, narrow, wide, thick, thin, slender, heavy, metal, metallic, wooden, plastic, glass, leather, denim, cotton, woolen, ceramic, stone, brick, concrete, paper, smooth, rough, soft, hard, fluffy, hairy, furry, shiny, glossy, matte, rusty, wet, dry, striped, checkered, patterned, spotted, dotted, printed, wearing, holding, carrying, sitting, standing, walking, running, smiling, lying, reclining, leaning, crouching, kneeling, jumping, riding, flying, swimming, eating, drinking, looking, pointing, reading, sleeping, driving, hanging, attached, mounted, folded, elderly, young, old, new, broken, dirty, clean, damaged, stained, burnt, full, empty, open, closed, bare, decorated, painted, covered, single, double, pair, couple, group, crowd, several, few, many

A qualitative observation of references produced by Molmo without any additional training was that many of them started by restating the prompt by mentioning the target region identifier — in this case, the ”red box”. This is a clear indicator that the models are not pragmatic, as they can not distinguish whether this information is useful for a listener. We investigate which reward setups successfully decrease the usage of such prefixes in Fig 9.

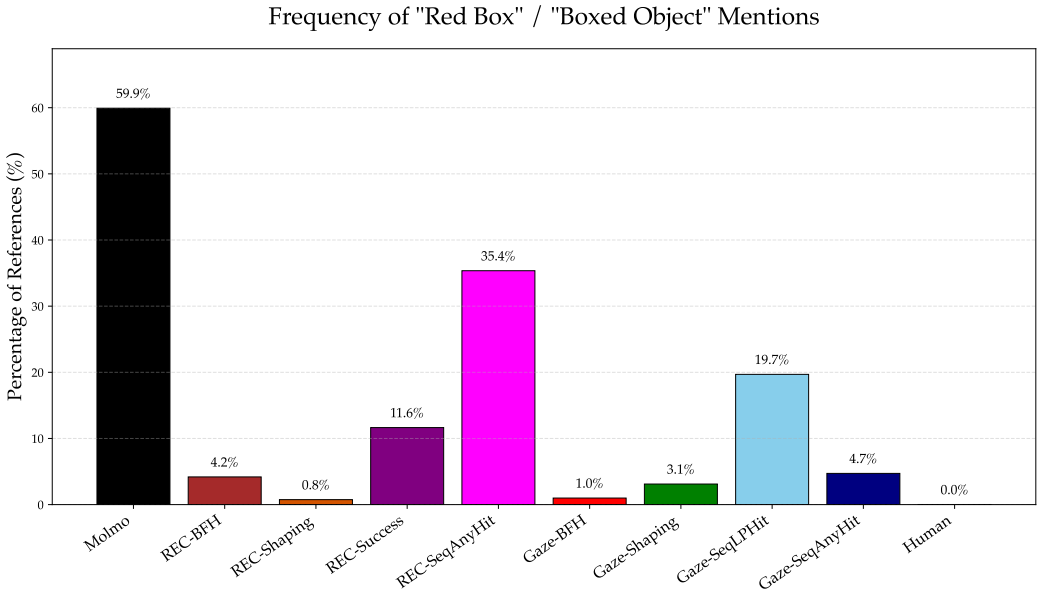


Figure 9: Mention frequency of the red bounding box across speaker models.

D.2 Reference quality annotations

We perform a small scale analysis of the quality and ambiguity of references. On a subset of reward configurations, one author annotated a subset of 50 image-reference pairs, sampled from each dataset as either well-specified, ambiguous, or incorrect. Configurations that encourage longer references such as REC-SUCCESS and GAZE-SEQANYHIT were more

likely to be well-specified. GAZE-SHAPING encourages shorter references, but also produces more ambiguous references, demonstrating the tradeoff between accuracy and efficiency.

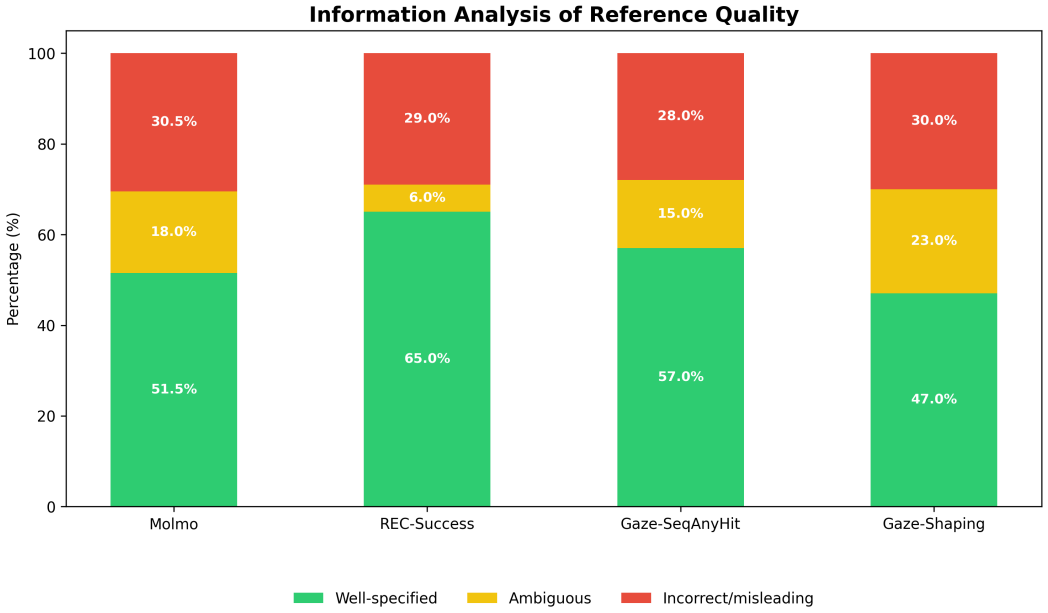


Figure 10: Analysis of ambiguity and incorrect reference rates across four selected speaker models.

D.3 Qualitative behavior and error discussion

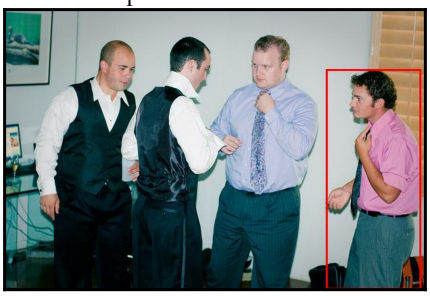
D.4 Gaze retention analysis

GAZE-BFH is the best performing model with machine listener, but it falls short with human listeners. Thus, we investigate why this might be the case. While not displayed in the examples in Fig 11, we suspect the GAZE-BFH configuration underperforms with human listeners due to its tendency to reward false positives. To corroborate this hypothesis, we analyze the post-hit target retention rate across speaker models, defined as the conditional probability that a gaze trajectory (or sequence of REC points) remains within the target bounding box after initial entry. Analysis results are shown in Table 13. Specifically, we evaluate the generated point sequences from the validation episodes of each model's best checkpoint (the checkpoint selected for evaluation). Overall, the gaze predictor is more prone to exiting the target region than the REC model (Molmo) since Molmo is trained for REC rather than realistic scanpaths. GAZE-BFH displays the worst retention rate of all speaker models. The reward setup does not encourage references that keep the listener's gaze in the bounding box. GAZE-SEQANYHIT also displays very poor retention rate, but likely does not suffer during human evaluation due to longer generations including more information.

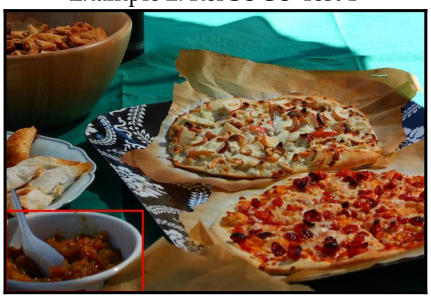
D.5 Punctuation and spacing errors

We observe that some of the configurations result in list-like references of many short sentences (e.g., GAZE-SEQLPHIT in Fig 11). Additionally, REC-SHAPING inserts extra spaces. To quantify these phenomena, we record the double-space rate of each speaker model along with the average number of periods in each reference. The latter metric is based on the assumption that most references are single sentences or phrases. In Table 14, GAZE-SEQLPHIT, GAZE-SEQANYHIT, and REC-SEQANYHIT are particularly prone to inserting extra periods. Only REC-SHAPING develops the double space behavior; sometimes, the

Example 1: RefCOCO Test A



Example 2: RefCOCO Test B



Speaker	Reference: Example 1	Reference: Example 2
Molmo	Man in pink shirt with short brown hair, looking left.	A white bowl with a red border containing a brownish-red sauce, located in the bottom left corner of the image.
REC-SeqAnyHit	A man wearing a pink shirt and gray pants, standing on the far right, with his left hand raised to his chin.	In the bottom left corner, there's a small white bowl containing a red sauce with a white spoon inside. This bowl is highlighted by a red box, making it stand out from the rest of the image.
REC-Success	Man in pink shirt, short brown hair, looking left, right hand raised to chin.	A white bowl containing a brownish-red sauce, located in the bottom left corner of the image, with a white spoon inside.
REC-Shaping	Man in pink shirt.	Bottom-left corner, white bowl with red border, containing brownish-red sauce.
REC-BFH	Man	A bowl of chili
Gaze-SeqAnyHit	Man wearing pink shirt and gray pants.	Bottom-left corner, white bowl with red border containing brownish-orange sauce.
Gaze-SeqLPHit	Dark-haired man. Pink shirt. Red-bordered box.	Bottom-left corner, white bowl with red outline, containing brownish-red sauce, spoon inside.
Gaze-Shaping	Man in pink shirt.	Bottom-left
Gaze-BFH	Man in pink shirt	White bowl with red sauce
Human	pink shirt	bowl on bottom left

Figure 11: Examples of generated references from all speaker models.

speaker generates spaces between tokens that make up a single word, affecting the speech synthesis during human evaluation. This exemplifies how the human evaluation accounts for behaviors that are otherwise difficult to quantify. We hypothesize both of these behaviors develop to exploit the listener model by giving it extra chances to resolve the reference through adding extra tokens without any additional semantic value.

Speaker Model	Gaze Retention Rate (%)
REC-SUCCESS	100.0%
REC-BFH	90.6%
REC-SHAPING	90.4%
REC-SEQANYHIT	85.7%
GAZE-SHAPING	54.3%
GAZE-SEQLPHIT	52.8%
GAZE-SEQANYHIT	39.9%
GAZE-BFH	37.2%

Table 13: Gaze trajectory retention metrics across speaker model configurations, demonstrating post-hit retention across reward setups.

Speaker Model	Mid-Ref Period Rate (%) ↓	Double Space Rate (%) ↓	Mean Periods
MOLMO	18.58	0.00	1.22
REC-BEFOREFIRSTHIT	0.78	0.00	0.02
REC-SHAPING	6.93	53.17	0.84
REC-SUCCESS	17.12	0.00	1.20
REC-SEQANYHIT	39.31	0.00	1.46
GAZE-BEFOREFIRSTHIT	0.52	0.00	0.04
GAZE-SHAPING	4.67	0.00	0.42
GAZE-SEQLPHIT	53.54	0.00	2.87
GAZE-SEQANYHIT	28.57	0.00	1.82
HUMAN	0.46	0.32	0.02

Table 14: Frequency of mid-reference period punctuation, double spacing, and mean period counts across speaker model configurations.

D.6 Coordinate generation

Finally, we occasionally observe references that are simply coordinates in REC-BFH (Table 15). Although not incredibly prevalent, this case highlights how using non human-like listeners can result in reward-hacking by creating references that exploit the machine listener.

Speaker Model	Coordinate Occurrence Rate (%) ↓
MOLMO	0.00%
REC-BEFOREFIRSTHIT	**0.78%**
REC-SHAPING	0.02%
REC-SUCCESS	0.00%
REC-SEQANYHIT	0.01%
GAZE-BEFOREFIRSTHIT	0.00%
GAZE-SHAPING	0.03%
GAZE-SEQLPHIT	0.01%
GAZE-SEQANYHIT	0.01%
HUMAN	0.00%

Table 15: Frequency of coordinate artifact occurrences across speaker model outputs.

E Human evaluation details

We recruit 751 participants on Prolific who are fluent English speakers in the United States. Participants receive $2 USD per session, each of which takes an average of 6.1 minutes. In each session, a participant completes up to 40 trials, with each trial corresponding to an image paired with a referring expression. We record all mouse movements, including clicks; only after the audio has finished playing will a click on the image move the participant to the next trial. Five of the 40 trials include human-written referring expressions; participants are provided a bonus of $0.50 USD if they successfully identify the target referent for all 5

human-written references without revealing the reference text. This bonus structure both incentivizes participants to perform well as a listener, and discourages them from revealing the reference text when unnecessary. To reduce spam and low-effort sessions, we discard all data from participants who do not successfully identify at least 4 of the 5 human-written references, resulting in a final pool of 633 participants.

E.1 Experimental instructions

The following instructions were provided to Prolific participants prior to beginning the reference game:

Experimental Instructions (Verbatim)

Welcome! This task consists of 40 or less trials. When each trial begins, you will see an image, and hear a description read aloud. Your job is to click on the region of the image that best matches the description.

Bonus! We will check your answers on a subset of trials with unambiguous descriptions. Participants who get 100% on this subset will receive a $0.50 bonus.

Need to see the description? If the speech description is unclear, you have the option to reveal the description text with a button below the image. This button will only appear after the speech description finishes. **However**, if you reveal the description for examples from the unambiguous subset, you will not receive the bonus, so only reveal it when you truly need it.

What if you don't know? There will be some ambiguous descriptions. If you are unsure, just make your best guess!

Ok, ready to begin? Please make sure your volume is turned up. The task will begin on the next page.