

Referring Expression Generation in Visually Grounded Dialogue with Discourse-aware Comprehension Guiding

Bram Willemsen and Gabriel Skantze

Division of Speech, Music and Hearing

KTH Royal Institute of Technology

Stockholm, Sweden

{bramw, skantze}@kth.se

Abstract

We propose an approach to referring expression generation (REG) in visually grounded dialogue that is meant to produce referring expressions (REs) that are both discriminative and discourse-appropriate. Our method constitutes a two-stage process. First, we model REG as a text- and image-conditioned next-token prediction task. REs are autoregressively generated based on their preceding linguistic context and a visual representation of the referent. Second, we propose the use of discourse-aware comprehension guiding as part of a generate-and-rerank strategy through which candidate REs generated with our REG model are reranked based on their discourse-dependent discriminatory power. Results from our human evaluation indicate that our proposed two-stage approach is effective in producing discriminative REs, with higher performance in terms of text-image retrieval accuracy for reranked REs compared to those generated using greedy decoding.

1 Introduction

A visually grounded dialogue is a conversation in which speakers refer to entities in a (shared) visual context. They do so by producing *referring expressions* (REs). The listener is expected to use the RE to identify the target entity, i.e., the *referent*. Whether the listener is successful in doing so depends on several factors, one being how specific the description of the referent was. With regard to specification, there exists a trade-off between discriminatory power and efficiency. On the one hand, the aim is to produce an unambiguous expression with which a referent can be successfully identified, whereas on the other hand a cooperative speaker is expected to make their contribution as economical as possible, while still avoiding ambiguity (Grice, 1975). To illustrate, consider the three phones depicted in Figure 1. If the intention of a speaker was to produce a description based on visual content that uniquely identified the phone second from



A: lets rank **the nokia No3** [...]

B: sorry, I don't get which nokia you are talking about.
the grey one or **the black one**?

Figure 1: Excerpt (simplified) taken from a dialogue collected by Willemsen et al. (2022).

the left, “*the phone with the QWERTY keyboard*” would be underspecified, as it applies to both the intended target as well as the right-most image. To avoid underspecification, additional content could be added to the RE, possibly resulting in a description such as “*the mostly black Nokia E75 mobile phone with the side-sliding QWERTY keyboard and keypad*”. This RE does set apart the target from the distractors, but is overspecified, as the description contains more content than is strictly required for identification of the referent in this context, violating the Gricean maxim of quantity (Grice, 1975).

In determining form and lexical content of REs, context plays a crucial role. We will again use Figure 1 to illustrate this by example. **A** attempts to draw the attention of **B** to a specific phone by referencing its brand name. However, since **B** recognizes two phones to be from this brand, **B** asks a clarification question that focuses on color. There are two things to note here. First, the REs produced by **B**, in particular “*the black one*”, only work as discriminative references due to the mention of the brand name just prior, as “*one*” is here a proform of “*nokia*” (the right-most phone is also black). Second is the symmetry between the REs, showing conventional preservation of form.

For a conversational agent to take part in visually grounded dialogue, it would preferably generate REs in a similar, context-dependent manner, as this is expected by human conversational partners. The computational modeling of this process is the do-

main of referring expression generation (REG), a core natural language generation (NLG) task for which a considerable body of work exists, spanning decades (see e.g., [Krahmer and van Deemter, 2019](#)). However, REG has traditionally focused primarily on the discriminative properties of REs, leaving discourse-appropriateness in the context of conversation a somewhat understudied problem.

In this paper, we propose an approach to REG for visually grounded dialogue that is meant to satisfy the discriminative property, while simultaneously accounting for discourse-appropriateness. We frame the problem as a two-stage process: in the first stage, we model REG as a text- and image-conditioned next-token prediction task: given a dialogue history, i.e., a preceding linguistic context, and the image of a referent, we autoregressively generate an RE as a continuation of the existing linguistic context, using a fine-tuned vision-language model (VLM). While at this stage we expect to generate an RE that fits the dialogue context and is indicative of the target image, it is not necessarily discriminative with respect to distractors. We, therefore, propose to use comprehension guiding as part of a *generate-and-rerank* strategy (see e.g., [Luo and Shakhnarovich, 2017](#)) in stage two; our goal being to select an RE with discriminative properties. Crucially, we introduce *discourse-aware* comprehension guiding as a way to estimate the discriminatory power of candidate REs based on the dialogue context and incorporate this in the candidate selection process.

Our main contributions are as follows:

- We propose an approach to REG in visually grounded dialogue based on causal language modeling with multimodal conditioning and fine-tune a generative VLM, here IDEFICS ([Laurençon et al., 2023](#)), for this purpose;
- We show the potential of *discourse-aware* comprehension guiding using the CRDG framework ([Willemsen et al., 2023](#)) as part of a modular REG system, with a higher average text-image retrieval accuracy for candidates selected with our reranking schema compared to greedily generated REs according to our human evaluation;
- We release the discussed materials, including our LoRA ([Hu et al., 2022](#)) weights for IDEFICS¹.

¹<https://github.com/willemsenbram/>

2 Related work

REG, as most NLG tasks, has been subject to a paradigm shift over the years. Whereas earlier methods were mostly symbolic (e.g., [Appelt, 1985](#); [Dale and Reiter, 1995](#); [Krahmer and Theune, 2002](#)), most approaches proposed in more recent years are based on neural models (e.g., [Mao et al., 2016](#); [Luo and Shakhnarovich, 2017](#); [Panagiaris et al., 2021](#); [Sun et al., 2023](#)). Contemporary NLG research frequently incorporates large language models (LLMs), predominantly those that are Transformer-based ([Vaswani et al., 2017](#)). A common approach to modeling downstream NLG tasks is domain adaptation via transfer learning. This is typically achieved by fine-tuning a pre-trained LLM on a task-specific dataset.

Although the bulk of the computation for most downstream tasks has been delegated to the pre-training of the base model, fine-tuning may still require significant computational resources. To combat this issue, parameter-efficient fine-tuning methods have been proposed, such as Low-Rank Adaptation (LoRA, [Hu et al., 2022](#)). By freezing the pretrained model weights and instead training rank decomposition matrices that have been added to the dense layers of the network, LoRA manages to reduce the number of trainable parameters by several orders of magnitude, often without considerable adverse effects to downstream performance.

Aside from language, Transformers have shown promising results when it comes to modeling other modalities (e.g., [Dosovitskiy et al., 2021](#); [Radford et al., 2023](#)). Of particular interest here are multimodal models that combine vision and language. VLMs such as CLIP ([Radford et al., 2021](#)) have learned to jointly embed both modalities via contrastive pretraining objectives. Their learned representations have shown to be useful for discriminative downstream vision-language tasks, such as text-image retrieval (TIR). We will hereafter refer to these models as discriminative VLMs. Other VLMs such as Flamingo ([Alayrac et al., 2022](#)), BLIP-2 ([Li et al., 2023](#)), Kosmos-2 ([Peng et al., 2024](#)), LLaVA ([Liu et al., 2023](#)), and InternVL ([Chen et al., 2024](#)) have been introduced to address *generative* downstream tasks, such as image captioning and (multi-turn) visual-question answering. These generative VLMs, sometimes called multimodal LLMs (MLLMs), are able to autoregressively output text based on multimodal inputs,

reg-with-guiding, doi:10.5281/zenodo.13225148

as they are built on (pretrained) LLMs with some form of visual input conditioning. This makes them particularly useful for inherently multimodal text generation problems such as REG for visually grounded dialogue.

REG has been defined as a task that is chiefly concerned with identification (Reiter and Dale, 1997). As such, most work in this area emphasizes the discriminative properties of REs. The goal is to generate an expression with which a referent can be unambiguously identified. Whether a candidate RE possesses this property is context-dependent, where context represents a multi-faceted concept.

One facet is the visual context in which the referent is embedded, often together with entities that may be mistaken for the referent, i.e., distractors. Various strategies have been proposed to have neural models take into account the visual context and attempt to maximize discriminatory power of generated REs, including discriminative decoding (e.g., Schüz and Zarriß, 2021) and comprehension-guiding (e.g., Luo and Shakhnarovich, 2017). These methods typically incorporate some manner of scoring (partial) candidate REs on the basis of their alignment with pragmatic principles, either at inference time to guide decoding, or as part of a *generate-and-rerank* strategy, a commonly used approach for a variety of NLG problems (e.g., Andreas and Klein, 2016; Challa et al., 2019; Won et al., 2023). In the latter case, a REG model will generate a set of candidate REs which are reranked on the basis of their discriminatory power according to some referring expression comprehension (REC) model.

These strategies, however, tend to focus primarily on the generation of definite descriptions, disregarding other forms of REs such as pronouns, and do not fully consider the dialogue context in which the REs would be used. Earlier work on rule-based REG did address some context-sensitive aspects, such as the by Krahmer and Theune (2002) proposed extensions to the influential Incremental Algorithm (Dale and Reiter, 1995), which included reduced descriptions of subsequent mentions and pronominalization. More recent work that explicitly considered the linguistic context in addition to the visual context has instead attempted to generate discriminative referring *utterances* (Takmaz et al., 2020), under the assumption, however, that each utterance only mentions a single referent.

3 Method

In this work, we focus on generating REs conditioned on a multimodal dialogue context for referents that are represented by independent images. This setting bares some resemblance to that of discriminative image captioning (see e.g., Vedantam et al., 2017; Cohn-Gordon et al., 2018; Schüz et al., 2021). REG more commonly attempts to describe objects or entities, represented by bounding boxes or segmentation masks, in single images or scenes. Spatial relations frequently become part of distinguishing descriptions in such settings as a result. Our method, however, focuses instead on generating REs based on visual content in situations that have been specifically designed for this to be challenging. We leave extending the framework to incorporate spatial relations to future work.

3.1 Task description

For a given referent, which is represented by an image (or images), the aim is to generate an RE (1) with which the referent can be identified and (2) which is discourse-appropriate.

3.2 Proposed approach

Broadly speaking, we propose a framework that consists of two components, namely a REG model and a REC model. For a visualization of this framework, see Figure 2. We approach REG as a causal language modeling problem. More specifically, we use a generative VLM that has been pre-trained to handle arbitrarily interleaved sequences of text and images (Alayrac et al., 2022; Laurençon et al., 2023) in order to condition the autoregressive generation of REs on a preceding visio-linguistic context. For the experiments presented in this paper, the generative VLM we use is IDEFICS (Laurençon et al., 2023), an open-source implementation of Flamingo (Alayrac et al., 2022). By fine-tuning IDEFICS on visually grounded dialogue data, our aim is to satisfy the second constraint of the task, i.e., generating REs that are a good fit for the projected use context. In order to ensure the generated REs satisfy the first constraint, we evaluate their discriminatory power using a REC model. Crucially, as part of a *generate-and-rerank* strategy, we propose *discourse-aware* comprehension guiding. The motivation for the use of a *discourse-aware* REC model to score discriminatory power comes from the context-dependence of this property, as some REs will need to be resolved to their

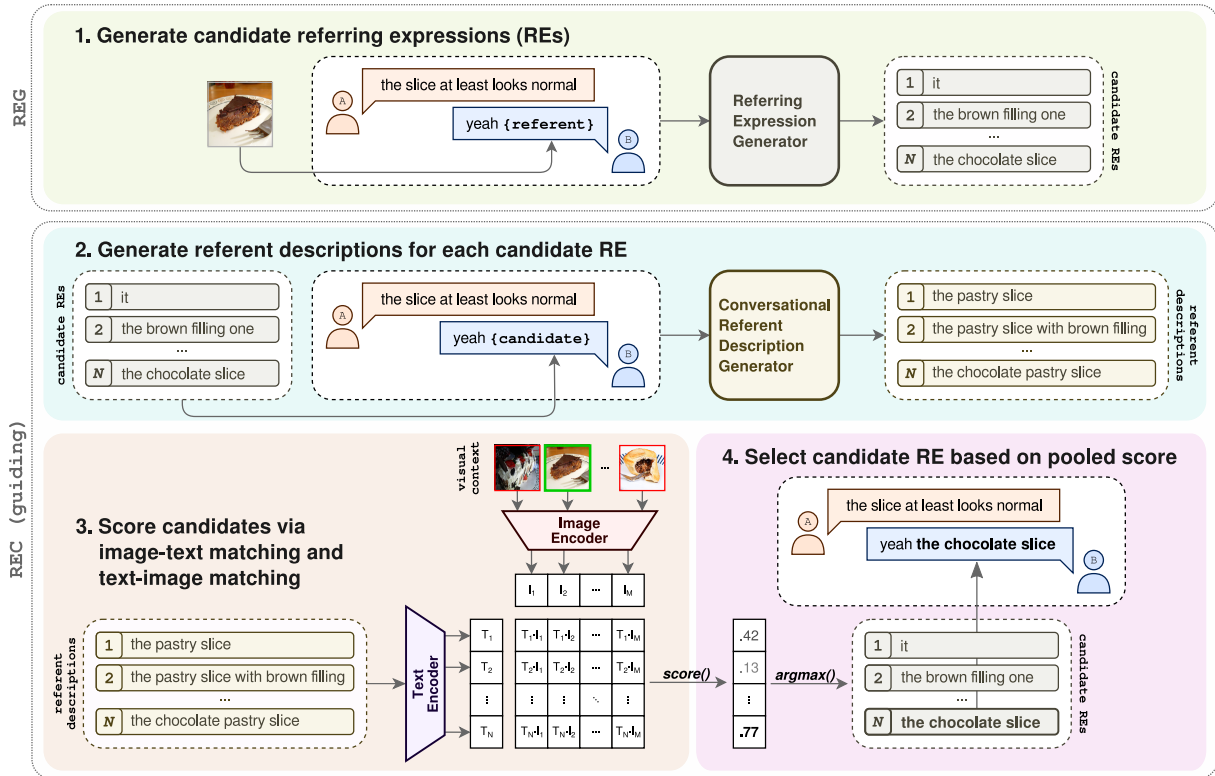


Figure 2: Visualization of the proposed two-stage, four-step framework. The first stage concerns (1) the autoregressive generation of candidate REs where the input to the REG model is the preceding linguistic context of the RE and an image representing the referent. In the second stage, candidate REs are (2) inserted into the dialogue segment at the point at which they were generated, after which the segment is processed by the CRDG (Willemsen et al., 2023) to generate referent descriptions. These referent descriptions are (3) used to evaluate the discourse-dependent discriminatory power of the candidate REs by using a pretrained VLM to produce TIM and ITM scores, which are then (4) weighted to arrive at a composite score for each candidate RE; the highest-scoring candidate RE is selected.

coreferences in order to be disambiguated and understood to be adequate mentions. For the experiments presented in this paper, we base our REC model on the conversational referent description generator (CRDG) framework of Willemsen et al. (2023).

3.2.1 Multimodal conditioning with IDEFICS

IDEFICS is a generative VLM based on the Flamingo VLM architecture (Alayrac et al., 2022). Flamingo was introduced to handle various opened vision-language tasks that carry an NLG objective, with a noted focus on using few-shot multimodal in-context learning (ICL) to accomplish them. Flamingo builds on pretrained vision and language models, bridging these modalities in order to incorporate visual information in the process of predicting the next token. To condition the autoregressive generation of text on both text and images, gated cross-attention dense layers that are trained from scratch are interleaved between the frozen layers of a pretrained LLM. Images are en-

coded using a pretrained vision model, after which the resulting embeddings go through a process of Perceiver-based (Jaegle et al., 2021) resampling in order to encode the high-dimensional visual feature representations as fixed numbers of so-called visual tokens. The model cross-attends to this output from the resampler in order to incorporate the visual information into its predictions, enabling the modeling of text interleaved with images.

To use IDEFICS for our purpose, we simply take the available linguistic context, indicating with speaker tokens the identity of the speaker for each message in the dialogue history, and add the image representing the referent to the sequence in the position at which we want to generate an RE. For reference, see step 1 in Figure 2.

3.2.2 Comprehension guiding with the CRDG

Willemsen et al. (2023) frame reference resolution in visually grounded dialogue as a TIR task. They note, however, that current discriminative VLMs, typically assume that the text is descriptive of the

image. As REs in dialogue can take various forms besides definite descriptions, being able to resolve coreferences, including pronouns, is often a prerequisite for successful identification of a referent. For this reason, they proposed fine-tuning a causal LLM to generate so-called *referent descriptions*. Referent descriptions distill all available coreferential information in the linguistic context of a given mention into a single (definite) description of the referent. These referent descriptions can then be used by a pretrained VLM to identify referents via (zero-shot) TIR. To illustrate, consider again the REs in Figure 1. If we were to attempt TIR directly with the RE “*the black one*”, the description is ambiguous, applying to both the target and a distractor. If we instead use its referent description “*the black nokia*”, which combines information from all mentions of the referent in the available linguistic context, we now have a distinguishing description. This shows how the linguistic context is crucially important in resolving an otherwise seemingly underspecified RE and how the CRDG can resolve references regardless of form.

While this framework was originally intended for REC in conversation, we propose to repurpose it as a comprehension-guiding model for REG in visually grounded dialogue. To evaluate candidate REs generated by our REG model based on their discriminatory power, we insert the candidate RE into the dialogue segment at the position at which it was generated by the REG model, marking its beginning and end in text. We then use the CRDG to autoregressively generate for this candidate RE a referent description based on the provided dialogue segment. For reference, see step 2 in Figure 2. The generated referent description is then encoded with a discriminative VLM to get a text embedding. We then compute representational similarity between this text embedding and the image embeddings of the candidate referents to rank the candidate REs. For reference, see step 3 in Figure 2. Note that the referent descriptions are only used in the process of guiding the selection of candidate REs.

Candidate reranking Although it makes intuitive sense to deem the candidate RE that has the most discriminatory power according to the REC model to be the best available candidate, this is not necessarily always true. To clarify, consider the following: if we were to simply opt for the candidate RE that has, among the candidates, the highest probability assigned to the target image via softmax, we may be selecting an RE based of a referent descrip-

TEXT-TEXT		TEXT-IMAGE	
Metric	Score	Metric	Score
BLEU	.71	Accuracy	.71
ROUGE-L	.82	MRR	.83
Jaccard	.79	NDCG	.88
Cosine _{TT}	.92	Cosine _{TI}	.48

Table 1: Cross-validated performance of incremental version of CRDG framework. Scores are rounded to the nearest hundredth.

tion that the VLM considers to be most similar to the target image when accounting for the distractors, but that is not in itself a good description of any of the images. Despite low similarity between the images and the description in absolute terms, the relative difference just so happens to be large and in favor of the target image. As a result, we would likely be selecting a suboptimal RE.

For this reason, we propose to select candidate REs not just based on their **text**→**image** matching (TIM) score, but rerank them based on both their TIM and **image**→**text** matching (ITM) scores: here, the TIM score indicates to what extent the candidate RE describes the target image with respect to the distractor images; the ITM score indicates to what extent the candidate RE describes the target image with respect to the other candidate REs. Note that each candidate RE is represented by its referent description, as generated by the CRDG, when these scores are computed. We combine the scores by way of linear opinion pooling (see e.g., [Jacobs, 1995](#)), taking a weighted linear combination of the log softmax of the TIM and ITM logits. For each candidate RE we calculate its pooled score, S , as follows:

$$S_i = w_{a_i} \cdot \ln(a_i + \varepsilon) + w_{b_i} \cdot \ln(b_i + \varepsilon)$$

where, for each i -th candidate RE, a and b represent its TIM and ITM softmax probabilities, respectively, each w the coefficient by which a and b are scaled, and ε a small constant that is added to avoid taking the (theoretical) log of 0. The coefficients sum to 1. We select the candidate RE with the highest S for the target image². We describe a hypothetical case in Appendix A to further illustrate the rationale behind this weighted reranking.

²Although we only consider the output from a single VLM here, it is possible to aggregate scores from multiple VLMs, treating each as an independent “expert”. Moreover, in addition to the VLM-based TIM and ITM scores, other properties of interest may also be incorporated as (weighted) “opinions”.

	TEXT-TEXT			TEXT-IMAGE			
	BLEU	ROUGE-L	Cosine _{TT}	Accuracy	MRR	NDCG	Cosine _{TI}
1-shot	.30	.34	.64	.57	.74	.80	.47
2-shot	.32	.36	.65	.58	.74	.81	.47
4-shot	.32	.35	.64	.53	.71	.78	.46
8-shot	.31	.34	.64	.49	.67	.76	.45
FT	.40	.48	.72	.67	.81	.86	.48

Table 2: Cross-validated n -shot and fine-tuned (FT) REG performance of IDEFICS using greedy decoding. Text generation metrics use *ground truth* REs as reference. Scores for TIR metrics are based on generated referent descriptions. Scores are rounded to the nearest hundredth.

4 Experiments

4.1 Data

The dialogues used in our experiments come from the visually grounded dialogue task A Game Of Sorts (AGOS, [Willemsen et al., 2022](#)). In this “game”, two players are presented with a set of nine images that they are asked to rank—one at a time—based on a given sorting criterion. To complete the task, they will have to agree on a ranking which they deem satisfactory. The game is played over multiple rounds with the same set of images to ensure repeated mentions of the same referents. Although the players see the same set of images, they cannot see each other’s perspective. The position of the nine images on screen is randomized, forcing the players to refer to the images based on their visual content. The task was specifically designed to encourage discussions and imposes no restrictions on message content. As a result, the referring language comes embedded in considerably longer and more diverse conversations compared to those from related work. [Willemsen et al. \(2022\)](#) collected 15 dialogues in total: three dialogues for each one of five image categories. Images from the same set were selected to have overlapping visual attributes, in order to further complicate the production of discriminative REs. Due to the deliberate challenges to the referential process and the relatively unconstrained nature of the dialogues, the task can be considered a challenging test bed for the grounding and generation of REs in conversation.

For fine-tuning and evaluation of both REG and REC models, we require dialogues with REs annotated. For this purpose, we use the span-based mention annotations for AGOS from [Willemsen et al. \(2023\)](#). These annotations indicate the start and end of all the mention spans found in the dialogues, and the image, or images, to which they refer. We will consider these human-produced REs

to be the *ground truth* for our study.

4.2 Evaluation

We focus on evaluating single-image referents, however noting that, in principle, our proposed framework can be extended to the multi-image referent case. We adopt the cross-validation protocol used by [Willemsen et al. \(2023\)](#), where the AGOS dataset is partitioned along the five image sets: for each run, twelve dialogues from four image sets are used for training, and the three dialogues of the remaining image set are used for testing. We limit the context window of the dialogue to the previous seven messages for model-based experiments, and report TIR results based on the reduced visual context, i.e., not considering ranked images to be part of the candidate referents.

4.2.1 Metrics

We score the referent descriptions generated by the CRDG based on their similarity to the manually constructed ground truth labels using the same metrics as reported in [Willemsen et al. \(2023\)](#), i.e., the Jaccard index, BLEU (based on unigrams and bigrams) ([Papineni et al., 2002](#)), ROUGE-L ([Lin, 2004](#)), and cosine similarity between text embeddings (Cosine_{TT}). When comparing generated REs against ground truth mentions, we compute unigram-based BLEU, ROUGE-L, and cosine similarity between text embeddings (Cosine_{TT})³. We report TIR performance in terms of top-1 accuracy, mean reciprocal rank (MRR), normalized discounted cumulative gain (NDCG), and cosine similarity between referent description text embeddings and target image embeddings (Cosine_{TI}). Model-based TIR results reflect the zero-shot performance of the discriminative VLM as it is used in the CRDG framework. This VLM is also used to

³Note that metrics based on overlapping content are not as robust for more open-ended tasks such as REG; we consider them here as secondary indicators for model selection.

	TEXT-TEXT			TEXT-IMAGE			
	BLEU	ROUGE-L	Cosine _{TT}	Accuracy	MRR	NDCG	Cosine _{TI}
Top-1	.21	.41	.71	.60	.76	.82	.47
Max disc.	.29	.40	.70	.89	.94	.95	.50
Rerank	.31	.40	.70	.86	.92	.94	.51

Table 3: Cross-validated REG performance of fine-tuned IDEFICS using beam search decoding with a width of 6. Text generation metrics use *ground truth* REs as reference. Scores for TIR metrics are based on generated referent descriptions. Scores are rounded to the nearest hundredth.

get the embeddings for the cosine similarity measures. All metrics are bound between $[0, 1]$.

4.2.2 Human

In order to externally validate our model-based experimental results, we conduct a human subjects experiment to evaluate human TIR performance for generated REs and to compare these results to those for the ground truth. Following Willemsen et al. (2023), participants are shown the REs in the context of the unfolding dialogue. We, however, show the dialogue up until the end of the current RE for which the participant is asked to provide an answer. We evaluate with the reduced visual context. For more details, see Appendix B.

4.3 Comparisons

Given the focus on multimodal ICL with Flamingo (Alayrac et al., 2022), we evaluate the n -shot performance of IDEFICS in addition to its (LoRA) fine-tuned performance. We compare these variants based on outputs generated using greedy decoding. For details about the selection of support examples for ICL, see Appendix C. Further experiments use the fine-tuned variants of the model. To generate multiple candidate REs, we use beam search with a width of 6. We examine how our proposed approach using weighted reranking (Rerank), which selects candidates based on their pooled score, compares against ablated versions of the method. We contrast performance with a variant that selects the candidate with the most discriminatory power (Max disc.) and a variant without any guiding that simply selects the top beam hypothesis (Top-1). We deliberately focus on evaluating different versions of the proposed framework, as, to the best of our knowledge, existing REG models are ill-suited to handle the AGOS task setting or principally do not satisfy our discourse-appropriateness criterion. For instance, if we were to use as a baseline a model that would invariably generate context-independent, but overspecified or caption-

like REs—such as discussed in Section 1 in relation to the example based around Figure 1—these may result in high TIR accuracy, but, even so, will virtually never be discourse-appropriate.

4.4 Implementation details

Similar to Willemsen et al. (2023), we obtain the CRDG by fine-tuning GPT-3—although davinci-002 instead of the davinci base model—using the OpenAI API. Crucially, however, our version of the CRDG is incremental as opposed to message-based. We use InternVL (Chen et al., 2024), specifically InternVL-G, as our discriminative VLM within the CRDG framework. With regard to the reranking of candidate REs, although we could treat the coefficients as learnable parameters, we instead simply set w to $\frac{2}{3}$ and $\frac{1}{3}$ for the TIM and ITM scores, respectively, as we believed this to represent a reasonable trade-off between the scores for our purpose. All experiments reported in this paper that involve IDEFICS are based on the 80 billion parameter variant⁴. We use quantized LoRA (QLoRA, Dettmers et al., 2023) for parameter-efficient fine-tuning. We modify the loss calculation by masking the loss for all tokens but the RE. We estimate, without exhaustive search, hyperparameters for IDEFICS fine-tuning using nested five-fold cross-validation. For additional details, including IDEFICS and GPT-3 hyperparameters, see Appendix D.

5 Results

Our results are based on 1305 of the 1319 annotated mentions of single-image referents; 14 samples were excluded as their target referents were not part of the set of candidate referents as a consequence of evaluating with the reduced visual context. Table 5 shows REs from different sources for a few dialogue samples.

Incremental CRDG Table 1 shows the performance of the CRDG on the ground truth data. We

⁴<https://huggingface.co/HuggingFaceM4/idefics-80b>

	Accuracy
Greedy	.74
Rerank	.78
Ground truth	.88

Table 4: Human (incremental) reference resolution performance. Scores are rounded to the nearest hundredth.

managed to closely replicate the results reported by [Willemssen et al. \(2023\)](#) despite our variant of the CRDG being incremental.

Multimodal ICL vs. fine-tuning In Table 2 we show results for candidate REs generated using greedy decoding with 1-, 2-, 4-, and 8-shot multimodal ICL and with the fine-tuned model. We found that a single example tended to be enough for the model to generate an RE, in accordance with the provided task. Adding an additional example improved performance slightly, but further increasing the number of support examples hurt performance instead. Moreover, the metrics showed a notable gap between ICL and fine-tuning, with fine-tuning averaging higher scores across the board.

Ablations Shown in Table 3 are results of the three strategies for candidate selection after beam search. With the exception of text-image cosine similarity, we observed slightly lower scores for the TIR metrics for the reranked REs in comparison with those that had the most discriminatory power. This was expected, as we actively went against taking the most discriminative candidate with our weighted reranking, which, our results suggested, did lead to higher representational similarity, on average, between referent descriptions and target images. These differences were, however, marginal.

Human performance We validated our model-based experimental results through human evaluation, results of which are shown in Table 4. We collected one data point per dialogue, meaning 15 data points per source of RE listed, for a total of 45 data points from 38 different participants. We contrasted TIR accuracy for REs generated with fine-tuned IDEFICS with that of ground truth mentions. We found that, although lagging behind the ground truth, the generated REs, regardless of the exact strategy, showed strong performance, far exceeding chance level (which was roughly 22%). Although both tested model-based RE variants seemed effective, our reranked REs resulted in higher accuracy than those based on greedy decoding.

RE length We found that REs generated by our (fine-tuned) REG model tend to be shorter, on av-

erage, than the ground truth mentions. This is one indicator of our model not having been prone to generating overspecified REs, which would otherwise have had the potential to artificially inflate accuracy scores. A comparison between the average length of the generated REs and the ground truth is visualized in Figure 4 in Appendix E.

RE content When examining the ground truth REs, we found that more than 20 percent of the included mentions contain no words that were descriptive of visual content (e.g., “it”, “that one”), with the pronoun “it” accounting for roughly half of these REs. We found that such REs were selected at a similar rate when using our weighted reranking schema. It is worth nothing, however, that whenever both the ground truth and selected candidate REs contained no content words, their forms would, at times, differ (e.g., “it” having been selected where the ground truth was “that one”).

6 Discussion

In this paper, we explored the problem of REG in visually grounded dialogue. Our aim was to realize the generation of REs that were not only discriminative, but also appropriate for the dialogue context in which they would be used. We proposed to approach the problem from a causal language modeling perspective, where the generation of tokens would be conditioned on both text and images. By fine-tuning a generative VLM, IDEFICS ([Laurençon et al., 2023](#)), we showed it is possible to generate REs that are indicative of the referent while suitable for the dialogue context. Notably, we were successful using a parameter-efficient fine-tuning approach ([Dettmers et al., 2023](#)) and while having relatively limited data for training ([Willemssen et al., 2022](#)). In addition, we introduced *discourse-aware* comprehension-guiding to evaluate whether candidate REs are discriminative given their linguistic context. By adding candidate REs to the dialogue for which they were generated, we were able to use the CRDG framework of [Willemssen et al. \(2023\)](#) to score candidate REs on their discourse-dependent discriminatory power. Finally, we showed that human TIR accuracy using candidate REs selected based on a weighted reranking of scores derived from this discourse-aware REC model was on average higher than for candidate REs generated through greedy decoding.

One of the main benefits of our approach is the ability for the REG model to generate REs that


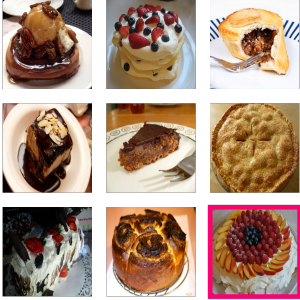

VISUAL CONTEXT			
LINGUISTIC CONTEXT	<p>[...]</p> <p>A: The poodle is the one that looks like a sheep right?</p> <p>B: yeah</p> <p>B: and now the husky</p> <p>A: Husky is {RE} right?</p>	<p>[...]</p> <p>A: the chocolate one now maybe? at least it has no cream, and some nuts</p> <p>B: ah true I didn't see the nuts there</p> <p>A: I'm not sure if it is ice cream to be honest</p> <p>B: The round one with lots of fruit? {RE}'s big and beautiful</p>	<p>[...]</p> <p>A: didnt we say the white SUV was more solid than grey and red?</p> <p>B: red then</p> <p>A: but sure we can swap</p> <p>A: {RE} now?</p>
<p>Greedy</p> <p>Top-1</p> <p>Max disc.</p> <p>Rerank</p> <p>GT</p>	<p>the one with the chain</p> <p>it</p> <p>it</p> <p>the one with the chain</p> <p>the one with a chain in the snow</p>	<p>It</p> <p>It</p> <p>It</p> <p>It</p> <p>It</p>	<p>white</p> <p>white</p> <p>white sedan</p> <p>white sedan</p> <p>white SUV</p>

Table 5: Examples of REs as produced by different versions of the proposed method, all generated with fine-tuned IDEFICS. **Greedy** shows REs generated using greedy decoding, **Top-1** means REs that were the top beam search result, **Max disc.** are REs generated with beam search that had the most discriminatory power, and **Rerank** are REs that were selected based on our weighted reranking. Also shown are the *ground truth* (**GT**) REs. The VISUAL CONTEXT depicts, for each dialogue, the unranked images at the time the ground truth RE was produced; the target referent is highlighted (magenta-colored border around the image). The LINGUISTIC CONTEXT shows (a limited number of) the preceding messages and the current message up until the start of the RE (**{RE}**); the light-gray text shows the remainder of the original message after the RE.

are commonly used in dialogue, but for which discriminatory power is neigh impossible to estimate without having an understanding of preceding linguistic context. A typical example of such REs are pronouns. As a result of our REC model being discourse-aware, our REG model is free to generate pronouns and other constructions involving proforms if these are deemed probable continuations of the current linguistic context, as the REC model will be able to evaluate whether these candidate REs are, in fact, discriminative.

With respect to the human evaluation, what is notable is that the model-based REs were generated based on a limited context window that included only the seven previous messages. The ground truth mentions, logically, were produced while the speakers had access to and knowledge of the entire dialogue history, the linguistic as well as the extralinguistic context. By evaluating using the unfolding dialogues in their entirety instead of limiting these to a rolling window of eight messages,

we biased the human evaluation slightly towards the ground truth; this was a conscious design choice as not doing so would unfavorably bias results towards the models instead. In light of this, our results are arguably even more promising.

Furthermore, rather than incorporating the entire visual context, our REG model was only conditioned on an image of the referent when generating an RE. As a result, the generated REs were generally descriptive, but not necessarily discriminative. Although we have now relied on our REC model to filter out such candidates, we suggest future research to consider the possibility of improving the generated candidates in terms of their discriminatory power by including the visual context as part of the input to the REG model. Related, we suggest testing alternative decoding strategies, for example those that are sampling-based or, perhaps more appropriate, ones that aim to be discriminative (e.g., Schüz and Zarriß, 2021).

Limitations

The experiments reported in this paper were based solely around modeling the English language; it is of yet unclear whether our results would transfer to other languages. We have focused on a single, relatively small dataset for which the annotations required by our approach were available; acquiring similar annotations for other, bigger datasets would be relatively costly. We have experimented with only one generative VLM for this paper; as a result, we do not know to what extent our findings generalize to other generative VLMs. We have used a closed-source API-based method for fine-tuning of the CRDG; consequently, we are not able to make the model weights publicly available, nor is the fine-tuning process transparent. The current iteration of the CRDG is unimodal, whereas the task of resolving references in visually grounded dialogue is inherently multimodal; this limits the maximally achievable performance. Our approach is modular and, as such, likely to be affected by error propagation; a bottleneck is the CRDG framework if it overvalues inadequate candidates (false positives) or undervalues adequate ones (false negatives) with respect to their discriminatory power. We currently operate on the assumption that utterance planning has been delegated to another system; this is a complex problem and challenging to solve properly, but will likely ultimately require a more unified approach that implicitly includes REG.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Dmytro Kalpakchi, Jim O'Regan, Travis Wiltshire, Chris Emmery, and the anonymous reviewers for their helpful comments.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. [Flamingo: a Visual Language Model](#)

[for Few-Shot Learning](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Jacob Andreas and Dan Klein. 2016. [Reasoning about Pragmatics with Neural Listeners and Speakers](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Douglas E. Appelt. 1985. [Planning english referring expressions](#). *Artificial Intelligence*, 26(1):1–33.

Ashwini Challa, Kartikeya Upasani, Anusha Balakrishnan, and Rajen Subba. 2019. [Generate, Filter, and Rank: Grammaticality Classification for Production-Ready NLG Systems](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Industry Papers)*, pages 214–225, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. [InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. [Pragmatically Informative Image Captioning with Character-Level Inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Robert Dale and Ehud Reiter. 1995. [Computational interpretations of the Gricean maxims in the generation of referring expressions](#). *Cognitive Science*, 19(2):233–263.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Fine-tuning of Quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. [An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale](#). In *The Ninth International Conference on Learning Representations (ICLR 2021)*. OpenReview.net.

Herbert Paul Grice. 1975. [Logic and Conversation](#). In Peter Cole and Jerry L. Morgan, editors, *Speech Acts*,

- volume 3 of *Syntax and Semantics*, pages 41–58. Academic Press, New York.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Robert A. Jacobs. 1995. [Methods For Combining Experts’ Probability Assessments](#). *Neural Computation*, 7(5):867–888.
- Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. [Perceiver: General Perception with Iterative Attention](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4651–4664. PMLR.
- Emiel Krahmer and Mariët Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing*, number 143 in CSLI lecture notes, pages 223–264. CSLI Publications.
- Emiel Krahmer and Kees van Deemter. 2019. [Computational Generation of Referring Expressions: An Updated Survey](#). In *The Oxford Handbook of Reference*. Oxford University Press.
- Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. [OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 71683–71702. Curran Associates, Inc.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.
- Chin-Yew Lin. 2004. [ROUGE: A Package for Automatic Evaluation of Summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual Instruction Tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Ruotian Luo and Gregory Shakhnarovich. 2017. [Comprehension-Guided Referring Expressions](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3125–3134.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. 2016. [Generation and Comprehension of Unambiguous Object Descriptions](#). In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11–20.
- Nikolaos Panagiaris, Emma Hart, and Dimitra Gkatzia. 2021. [Generating unambiguous and diverse referring expressions](#). *Computer Speech & Language*, 68:101184.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a Method for Automatic Evaluation of Machine Translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shao-han Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. [Grounding Multimodal Large Language Models to the World](#). In *The Twelfth International Conference on Learning Representations (ICLR 2024)*. OpenReview.net.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning Transferable Visual Models From Natural Language Supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. [Robust Speech Recognition via Large-Scale Weak Supervision](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.
- Ehud Reiter and Robert Dale. 1997. [Building applied natural language generation systems](#). *Natural Language Engineering*, 3(1):57–87.
- Simeon Schüz, Ting Han, and Sina Zarriß. 2021. [Diversity as a By-Product: Goal-oriented Language Generation Leads to Linguistic Variation](#). In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 411–422, Singapore and Online. Association for Computational Linguistics.
- Simeon Schüz and Sina Zarriß. 2021. [Decoupling Pragmatics: Discriminative Decoding for Referring Expression Generation](#). In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 47–52, Gothenburg, Sweden. Association for Computational Linguistics.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. 2023. [A Proposal-Free One-Stage Framework for Referring Expression Comprehension and Generation via Dense Cross-Attention](#). *IEEE Transactions on Multimedia*, 25:2446–2458.

Ece Takmaz, Mario Giulianelli, Sandro Pezzelle, Arabella Sinclair, and Raquel Fernández. 2020. [Refer, Reuse, Reduce: Generating Subsequent References in Visual and Conversational Contexts](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4350–4368, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. [Context-Aware Captions from Context-Agnostic Supervision](#). In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079.

Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. [Collecting Visually-Grounded Dialogue with A Game Of Sorts](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2257–2268, Marseille, France. European Language Resources Association.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. [Resolving References in Visually-Grounded Dialogue via Text Generation](#). In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–469, Prague, Czechia. Association for Computational Linguistics.

Seungpil Won, Heeyoung Kwak, Joongbo Shin, Janghoon Han, and Kyomin Jung. 2023. [BREAK: Breaking the Dialogue State Tracking Barrier with Beam Search and Re-ranking](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2832–2846, Toronto, Canada. Association for Computational Linguistics.

A Reranking

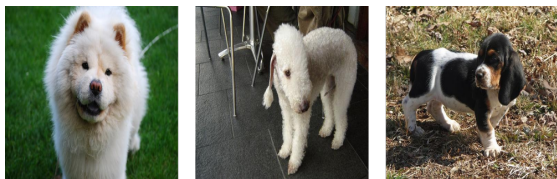


Figure 3: Images of dogs for the example in Appendix A to illustrate the rationale behind weighted reranking.

We will further illustrate the need for reranking using a simplified, hypothetical example based

around the images in Figure 3. Figure 3 depicts three images of dogs. We will consider the leftmost image to be our target, with the other two serving as distractors. We have three candidate REs for the target image: “*the white dog*”, “*the green car*”, and “*the attentive dog*”. Of these three candidates, “*the attentive dog*” is arguably the most appropriate. The RE “*the green car*” does not fit the target image nor does it describe the distractors, as none depict a car. The RE “*the white dog*” is underspecified, as it applies to both the target image and a distractor (the middle image). Given that the target image depicts a dog that looked directly at the camera when its picture was taken, which is not true for the other dogs, using the adjective “*attentive*” should be acceptable.

Now, in order to perform candidate selection, we use a discriminative VLM to encode each candidate RE and each image that is part of the visual context. If we then compute representational similarity between text and image embeddings, followed by a softmax over the resulting logits per candidate RE, we get what we consider a probability distribution over the images per candidate RE. This is expected to provide some indication with respect to how well the target image is described by each candidate RE given the current visual context.

However, in the scenario that we have sketched here, the following may happen. Although “*the green car*” has low representational similarity in absolute terms with each image, due to the greater presence of the color green in the target image it scores considerably higher than the distractor images for this candidate RE, which is amplified by the application of the softmax function. As a result, in this hypothetical, the softmax score for the target image for the candidate RE “*the green car*” would be considerably higher than the score of the more appropriate “*the attentive dog*”. Clearly, selecting REs based solely on this score is not appropriate.

One way to address this is to not only apply the softmax over the images per candidate RE, but to also apply it over the candidate REs for the target image. This will provide an indication for how well the target image is described by each candidate RE, in relation to the other candidates. The highest softmax score is likely assigned to “*the white dog*”, with “*the attentive dog*” in close second, and “*the green car*” a distant third. The candidate “*the white dog*” would be an acceptable RE were it not for the fact that it also applies to a distractor. If we were to select REs based solely on this score, we are more

likely to select a candidate that is descriptive, but not discriminative.

Thus, we instead combine the two scores to arrive at a composite that more accurately represents the appropriateness of the candidate REs in the given context than each score independently would. We gain further control over the trade-off between descriptive and discriminative through weighting.

B Human evaluation

Instructions provided to participants are shown in Figure 6 and Figure 7, with the informed consent question shown in Figure 8. An example of a task-related question is shown in Figure 5. The order of the images is randomized per question. An attention check is added after every 25 task-related questions. The survey platform we used was LimeSurvey⁵, with participants recruited via Prolific⁶. Eligible workers had a minimum approval rate of 99%, a minimum of 500 previously completed submissions, and had indicated that they are fluent in English. Regardless of the source of the RE, the participants were allowed to provide data for at most one dialogue per image set. The expected time-on-task was adjusted based on the number of questions, which varied due to a variable number of REs per dialogue. Participants were financially compensated for their contributions, with compensation affected by the expected time-on-task.

C Support examples

In order to select suitable support examples for multimodal ICL, we examined the dialogues to find the most frequently occurring forms of REs. We identified four categories of REs for which we selected two support examples per image category. The RE categories were (in)definite descriptions (e.g., “the white curly dog”), pronouns (e.g., “it”), noun phrases that included a proform in addition to content words (e.g., “the black one”), and noun phrases that contained no content words (e.g., “that one”). They are listed here in order of importance, meaning for 1-shot ICL the support example was taken from the (in)definite descriptions category, 2-shot had a support example for both the (in)definite descriptions and pronouns categories, and so on. For each support example we added the preceding seven messages from the dialogue history and the (partial) task description that was shown to the

⁵<https://www.limesurvey.org/>

⁶<https://www.prolific.com/>

participants. Examples were formatted according to the “User-Assistant” template, where the “User” provides the dialogue segment up until the start of the RE and the “Assistant” provides the RE in response.

D Additional implementation details

For both fine-tuning and inference, we distribute the model over 8 x 24GB NVIDIA GeForce RTX 3090 using naive model parallelism. Hyperparameters for IDEFICS fine-tuning are provided in Table 6. Hyperparameters for GPT-3 fine-tuning via the OpenAI API are provided in Table 7.

Training samples for IDEFICS fine-tuning were formatted as follows:

```
[bos token] +
[preceding linguistic context] +
[referent image] +
[start of RE token] +
[RE] +
[end of RE token] +
[eos token]
```

Note that the preceding linguistic context included a (partial) task description. Separate messages were joined by newline characters. The following is an example of a sample (shortened window for illustrative purposes):

```
<s> M: Your neighbour’s cat frequently uses your
garden as its own personal bathroom. You decide
to adopt a dog to deal with this issue. Which of
these dogs would be most effective in scaring off
the neighbour’s cat and why?\nA: yeah lets go
for chow\nB: And then <referent_image> >> the
husky << </s>
```

Epochs	1
Batch size	1
Gradient accumulation steps	4
Learning rate	7e-5
LoRA r	16
LoRA α	32
LoRA dropout	0.1

Table 6: Hyperparameters for fine-tuning of IDEFICS-80b. We use default values if not otherwise specified.

Epochs	3
Batch size	2
Learning rate multiplier	2

Table 7: Available hyperparameters for fine-tuning of GPT-3 (davinci-002) using the OpenAI API.

E Additional results

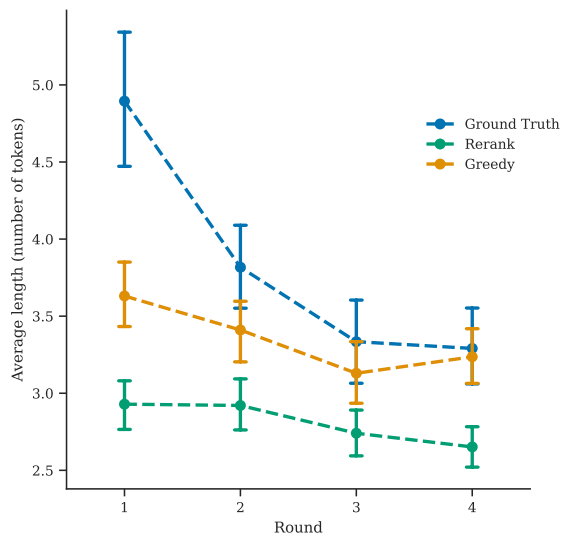


Figure 4: Average RE length per round. Shown are *ground truth* REs taken from the dialogues (blue), REs generated by the fine-tuned IDEFICS model using greedy decoding (orange), and REs selected based on our weighted reranking (green). Error bars indicate 95% bootstrapped confidence intervals.

Task: You are looking to hang a picture on your wall, but you have no hammer at your disposal to put the nail in the wall. Which of these phones would you consider most suitable to use as an impromptu hammer and why?

Please discuss the scenario and come to an agreement on how to rank these phones (starting with the phone that is most suitable) and motivate your choices!

A: Hello!

B: Hello!

B: How you doin'?

A: Good! How are you?

B: Great!

A: Ok! Let's start?

B: Yes!

A: On top of my head, I would not use an expensive phone as a temporary hammer, so maybe we can rank by how expensive the phone is?

B: Fair enough. The first thing I was thinking about was material en how thick the phones are.

B: And the amount of glass

A: Right, so we need a thick one with less glass?

B: I think so.

B: I think there all plastic

A: I think so too. So maybe we can't really distinguish by material.

A: We can take the nokia one with a protective casing? The glass seems small on that one.

B: Yes, and the one with the least of glass are perhaps also the cheapest ones

A: Agreed.

B: Yes i'd like that idea

A: Nice. So the next one for me is → the Samsung flip phone ←

*Which image did **A** refer to?
(between the → and ← arrows)

Choose one of the following answers

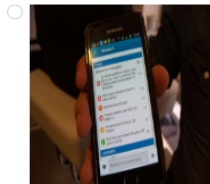
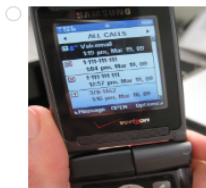
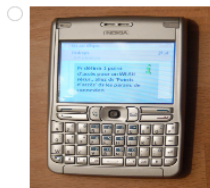


Figure 5: Example of an item shown to participants during the human evaluation study.

Instructions

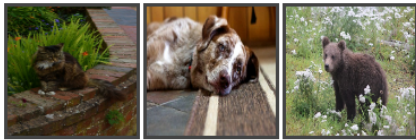
You will read a conversation between two people that are discussing images. The two people that are having this conversation are playing a game in which they have to rank 9 images. This game is played over 4 rounds. Each round the players are given a different scenario (indicated by "Task", in bold letters) by which they have to rank the images (for example: "which of these animals is the cutest?"). Each time you see a new scenario a new round has started.

We ask you to find which image the players are referring to each time they mention one of the images shown below the conversation. To help you see when and where a player mentioned an image, the expression for which we need you to find the associated image will be marked with two red arrows (→ and ←). This expression has either been produced by a human or by a machine. As you progress, you will be able to see increasingly more of the dialogue. Use that to your advantage, but note that you cannot undo any of your previous selections. In the event that the expression was machine-generated, you will be able to see the actual human-produced expression in its place as you continue with the task after having made your selection.

Note that this is a relatively long task! We estimate it to take **30 minutes on average**. We will need you to pay close attention all the way through to the end. But you will be compensated accordingly for your efforts! **Be aware that there will occasionally be an attention check.**

To illustrate the task, please have a look at the following [example](#):

A: I think → the dog ←

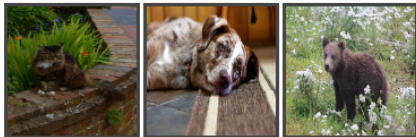


Which image did **A** refer to?
(between the → and ← arrows)

Answer: Here, "the dog" refers to the image of the dog, thus you would select the image in the middle

A: I think the dog is cute. What do you think?

B: I also think → it ←



Which image did **B** refer to?
(between the → and ← arrows)

Answer: Here, "it" refers back to "the dog", which refers to the image of the dog, thus you would again select the image in the middle

Figure 6: Instructions as shown to participants during the human evaluation study (1/2).

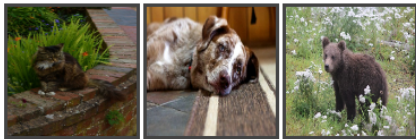
A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like → the cat ←



Which image did B refer to?
(between the → and ← arrows)

Answer: Here, "the cat" refers to the image of the cat, thus you would select the leftmost image

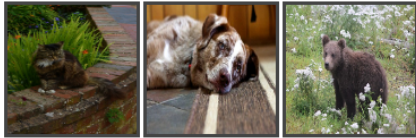
A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. → It ←



Which image did B refer to?
(between the → and ← arrows)

Answer: Here, "It" refers back to "the cat", which refers to the image of the cat, thus you would again select the leftmost image

A: I think the dog is cute. What do you think?
B: I also think it is cute.
B: I do like the cat as well. It looks a bit grumpy, but → it ←



Which image did B refer to?
(between the → and ← arrows)

Answer: Here, "it" again refers back to "the cat", which refers to the image of the cat, thus you would again select the leftmost image

On a final note, be aware that as players progressed through their game, they would rank images along the way. A ranked image is no longer shown as one of the possible images to select for the current round, but all images will again be available for selection at the start of a new round: the players are given a new scenario and the same set of images to rank at the start of each round.

Figure 7: Instructions as shown to participants during the human evaluation study (2/2).

*By clicking "Yes", you indicate that you have read the instructions and are willing to participate in this study.
The main purpose of this data collection is to understand to what extent people are capable of grounding referring expressions in conversation.
You agree that any data we collect from you through your participation in this survey may be used for research purposes and in publications.
Any such data will be anonymized prior to publication.
We will manage your data in accordance with the General Data Protection Regulation (GDPR).
This means that you have the right to withdraw your consent, request your data, and request that your data be deleted, at any time.
Your participation in this study is voluntary and you may decide to stop at any point.
Note that not completing the survey will affect your compensation.
In case you have any questions or concerns you can contact us by sending a message on Prolific or by sending an email to [redacted]

Yes No

Figure 8: Participant informed consent for human evaluation study.