Resolving References in Visually-Grounded Dialogue via Text Generation

Anonymous ACL submission

Abstract

Vision-language models (VLMs) have shown to be effective at image retrieval based on simple text queries, but text-image retrieval based on conversational input remains a challenge. Consequently, if we want to use VLMs for reference resolution in visually-grounded dialogue, the discourse processing capabilities of these models need to be augmented. To address this issue, we propose fine-tuning a causal large language model (LLM) to generate definite descriptions that summarize coreferential information found in the linguistic context of references. We then use a pretrained VLM to identify referents based on the generated descriptions, zero-shot. We evaluate our approach on a manually annotated dataset of visually-grounded dialogues and achieve results that, on average, exceed the performance of the baselines we compare against. Furthermore, we find that using referent descriptions based on larger context windows has the potential to yield higher returns.

1 Introduction

Visually-grounded dialogues are conversations in which participants make references to the visual world. Referring in conversation is understood to be a collaborative process, with shared responsibility for ensuring the successful identification of the referent (Clark and Wilkes-Gibbs, 1986). It is not uncommon for a definite reference to be established over multiple turns, with each separate contribution unlikely to be a minimally distinguishable description of the referent. Taken out of their use context, these referring expressions may be difficult, if not impossible, to resolve. Consider the example dialogue in Figure 1. The underspecified description *"the shiny one"* leads to a clarification question, *"Do you mean that red one?"*. To resolve the expression *"that red one"* to its referent, we need information from earlier in the conversation to understand that *"one"* is a proform of *"apple"*.

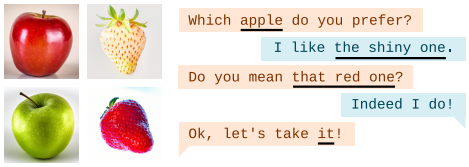

Figure 1: Example dialogue in which two participants discuss fruits. Expressions that denote one or more images are underlined.

Without this linguistic context, the red strawberry and the red apple are equally likely referents.

We can break the problem of reference resolution in visually-grounded dialogue down into three subproblems: (1) mention detection, or finding the expressions that can be grounded in the visual context (*"that red one"*); (2) aggregation of referent-specific information (linking *"apple"*, *"the shiny one"*, and *"that red one"*); and (3) referent identification, or the grounding of language (finding the referent that is best described by the three expressions from among a set of candidate referents). This final step requires bridging the gap between vision and language. For this purpose, we can turn to pretrained vision-language models (VLMs), which have shown to be effective at zero-shot text-image retrieval when given a description of an image (e.g., Radford et al., 2021; Jia et al., 2021; Li et al., 2023). However, current VLMs lack the discourse processing capabilities necessary for reference resolution in visually-grounded dialogue. Although some VLMs may correctly identify the red apple as the referent given the entire dialogue of Figure 1, dialogues are often vastly more complex than this hypothetical exchange. Take, for instance, the dialogue in Appendix A: with multiple mentions of different referents within the same utterance, such a brute-force method would immediately fail. It is clear that if we want VLMs to be effective for this purpose, their discourse processing capabilities need to be augmented.

To this end, we propose fine-tuning a causal large

language model (LLM) for the task of *referent description generation*. Referent description generation can be regarded as a special case of referring expression generation with the goal of always generating the most complete expression possible. For a given mention, the model is trained to generate a definite description that summarizes all information that has been explicitly disclosed about the referent during a conversation. For example, for the mention *"that red one"* in Figure 1 we would want the model to generate the description *"the shiny red apple"*. We will refer to the fine-tuned model as the *conversational referent description generator* (CRDG). The description generated by the CRDG is then used by a pretrained VLM to identify the referent, zero-shot. Our approach can be seen as an exploration of the limits of depending on linguistic context alone for generating referent descriptions, as the discourse processing and eventual grounding of the descriptions are entirely disjoint.

For the experiments presented in this paper we use data from the collaborative image ranking task A Game Of Sorts (AGOS, Willemsen et al., 2022). Referents are represented by separate, but visually similar images from a shared entity category. Due to their largely unrestricted nature and with a focus on the collaborative referential process, the collected dialogues form a challenging test bed for visually-grounded language understanding in conversation. We manually annotate the dialogues by marking mention spans and aligning the spans with the images they denote, and provide ground truth referent descriptions for all marked mentions.

Our main contributions are as follows:

- We present a generative approach to reference resolution in visually-grounded dialogue that frames the discourse processing side of the task as a causal language modeling problem;
- We show that it is possible to fine-tune a causal LLM to generate referent descriptions from dialogue to be used by a pretrained VLM for referent identification, zero-shot;
- We release the discussed materials, including our annotations for A Game Of Sorts (Willemsen et al., 2022)[1].

2 Background

Visually-grounded language understanding is fundamental for conversational agents that engage in dialogue involving references to the visual world. Researchers have introduced a variety of tasks that provide data for development and frameworks for evaluation of visually-grounded dialogue models. These tasks often take the form of goal-oriented, dyadic interactions but differ in terms of, for example, the visual stimuli used, e.g. abstract figures or realistic photos; the roles assigned to participants, e.g. whether symmetric or asymmetric; constraints on message content, e.g. a fixed vocabulary; and the nature of the task, e.g. navigation, identification, ranking, and multi-turn visual question answering (e.g. Das et al., 2017; De Vries et al., 2017; Shore et al., 2018; Ilinykh et al., 2019; Haber et al., 2019; Udagawa and Aizawa, 2019; Willemsen et al., 2022). It has been noted that the task configuration can significantly impact the extent to which certain dialogue phenomena, such as coreferences and clarification requests, are represented in the collected data, if at all (Agarwal et al., 2020; Haber et al., 2019; Ilinykh et al., 2019; Schlangen, 2019; Willemsen et al., 2022). Tasks that heavily constrain the interactions do not reflect the complex nature of dialogue to the same degree as tasks that have been designed for these phenomena to naturally emerge as part of the discourse, such as A Game Of Sorts (Willemsen et al., 2022), which we use in this paper.

The terms referring expression comprehension (e.g. Yu et al., 2016), referring expression grounding (e.g. Zhang et al., 2018), referring expression recognition (e.g. Cirik et al., 2018), and reference resolution (e.g. Kennington et al., 2015) have been used interchangeably to describe the problem of mapping the language that denotes a referent to a representation of that referent in the visual modality. Prior work noted the importance of referring expressions to conversation, but often modeled the problem independent of the dialogue (e.g. Cirik et al., 2018; Schlangen et al., 2016; Yu et al., 2016; Zhang et al., 2018). The granularity at which grounding occurs may differ between works, as the language may be mapped to bounding boxes of individual objects (Cirik et al., 2018; Schlangen et al., 2016; Yu et al., 2016; Zhang et al., 2018), objects or larger image regions represented by segmentation masks (Liu et al., 2017), or entire images altogether (Haber et al., 2019; Takmaz et al., 2020).

To address the problem computationally, both modalities must in some way be encoded. Engineered visual feature representations and sim-

[1] Available upon publication.

ple language models such as those based on n-grams (e.g. Kennington et al., 2015; Kennington and Schlangen, 2017; Shore and Skantze, 2018) have been mostly replaced with more powerful learned representations that embed the images and text in high-dimensional vector spaces (Haber et al., 2019; Takmaz et al., 2020). This has made it possible to resolve references by computing representational similarity between an encoding of the text that contains a mention and the embeddings of the candidate referents, where the candidate that has the highest matching score is assumed to be the referent (Haber et al., 2019; Takmaz et al., 2020).

Recent work on multimodal representation learning has shown that jointly embedding text and images can work at scale. Trained using a contrastive objective, maximizing representational similarity between true pairings of images and text while simultaneously minimizing similarity of false pairs, vision-language models (VLMs) such as CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022), and BLIP-2 (Li et al., 2023), have shown to be effective zero-shot classifiers, outperforming the previous state-of-the-art on various benchmarks without the need for further fine-tuning on specific tasks. However, despite their noteworthy image-text matching performance based on simple text queries, these VLMs lack the discourse processing capabilities required for reference resolution in visually-grounded dialogue. Even a simplified example, such as shown in Figure 1, illustrates a fundamental challenge, namely that of coreference resolution. The interpretation of anaphoric pronouns such as *"it"* is dependent on their antecedents. Without resolving its coreferences first, identifying the referent based on the pronoun alone leads to a random guess.

To improve downstream performance on discourse processing tasks involving coreference, prior work has approached the problem as one of transforming the original input based on linguistic context. This was done either via substitution, such as in Bhattacharjee et al. (2020) where pronouns were substituted for more descriptive mentions of the same referent, or via generation, such as in Quan et al. (2019) where entire utterances were reconstructed in a pragmatically complete manner with coreferences and ellipses resolved. To the best of our knowledge, this approach has not yet been applied to reference resolution in visually-grounded dialogue.

Most contemporary natural language processing (NLP) works use Transformer-based language models (Vaswani et al., 2017). For text generation tasks, it is common to use (unidirectional) autoregressive, or *causal*, language models such as GPT (Radford et al., 2018). While processing sequences, causal language models mask the future, allowing the model to only attend to the current and previous tokens while predicting the next token. A persistent trend has been to scale up language models, both in terms of their parameter count and the size of their training datasets. These increasingly larger models, such as GPT-3 (Brown et al., 2020), OPT (Zhang et al., 2022), PaLM (Chowdhery et al., 2022), and LLaMa (Touvron et al., 2023), have been dubbed *large language models* (LLMs). The current leading paradigm to modeling downstream NLP tasks is based on transfer learning, where a pretrained LLM is fine-tuned for a specific task on a smaller, domain-specific dataset.

3 Method

We treat visually-grounded reference resolution as a text-image retrieval task, where referents are represented by images. We leave finer-grained grounding of words and phrases to image regions or individual entities or parts thereof for future work.

3.1 Proposed Framework

We frame the discourse processing side of the task as a causal language modeling problem. Figure 2 shows a visualization of the proposed framework.

Task Definition We denote the dialogue as $D = (u_1, u_2, ..., u_n)$, where each u_i represents an utterance. Each utterance consists of an ordered sequence of tokens. An utterance may contain one or more mentions, denoted as M. A mention is an ordered subsequence of tokens from an utterance. A mention has an exophoric referent, denoted as R. A mention is embedded in what we call its linguistic context, denoted as L. As an ordered subsequence of D, the linguistic context of a given mention consists of the utterance in which it is contained and all preceding utterances. The number of preceding utterances, hereafter referred to as the dialogue history, may be capped if a finite size context window is defined. The aim of visually-grounded reference resolution is to resolve a reference to its referent, i.e. to identify R for a given M, from a set of candidate referents, denoted as C, such that $R \subseteq C$.

Referent Description Generation We propose to

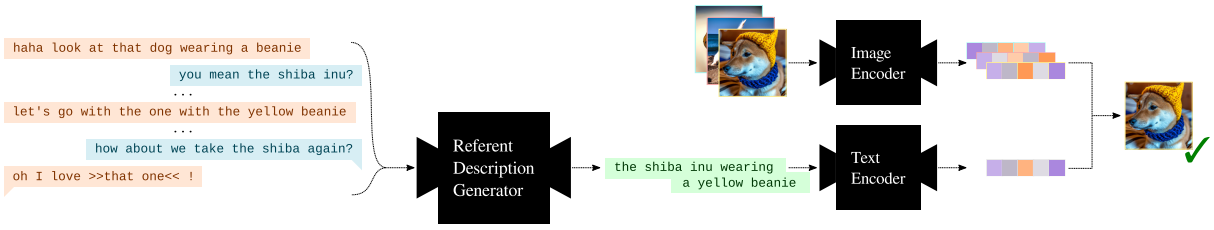

Figure 2: Visualization of the proposed visually-grounded reference resolution framework. With the CRDG we generate a referent description for a marked mention, to be used by a (frozen) pretrained VLM for referent identification.

generate a definite description, denoted as Y, for a given mention M that summarizes all that has been disclosed in L about the referent R. For this purpose, we fine-tune a causal LLM that learns to generate Y conditioned on L. Y is a sequence of tokens expected to be largely constructed from tokens that appear, or are some derivative of tokens that appear, in the coreference chain of R, which is contained in L. We refer to the fine-tuned model as the *conversational referent description generator* (CRDG). For an example of the context dependency of referent description content, see Figure 4 in Appendix B.

LLM Input We mark M in u_i by inserting positional markers as special tokens to indicate the beginning and end of the mention span. We prepend each utterance in L with a speaker token to indicate the source of the contribution. When D is task-oriented, we update L by prepending task instructions, i.e. a special token followed by a sequence of tokens describing the task performed by the dialogue participants. For an example of the input to the LLM, see Figure 5 in Appendix B.

Text-Image Retrieval We use a pretained VLM to identify R from C based on Y, zero-shot. We use the text encoder of the VLM to encode Y into an n-dimensional feature vector, denoted as $\mathbf{v}$. We use the image encoder of the VLM to encode each candidate referent of C into an n-dimensional feature vector, which gives a $|C| \times n$ matrix, denoted as $\mathbf{A}$. We then compute their matrix-vector product. For single-image referents, i.e. when $|R| = 1$, we take the referent to be $R = argmax(\mathbf{Av})$.

In order to produce accurate referent descriptions, the CRDG must implicitly learn to perform coreference resolution as we do not provide explicit supervision for this subtask. In each sample, only the current mention for which we want the model to generate a description is marked; none of its coreferences are in any way indicated. A principal advantage of our model is that it can resolve multiple mentions, even when they have different referents, appearing in the same utterance, including nested mentions. Note that for the purpose of this study, we assume mention detection to be solved. As it stands, using this framework in production requires a separate model to propose candidate mentions at the span level.

3.2 Baseline Models

As a lower bound, we report random chance performance. In addition, we compare performance of our approach to baselines based on simple heuristics and a coreference resolution model.

3.2.1 Heuristics

Mention We evaluate the image retrieval performance when the VLMs are presented with just the marked mentions.

Substitution We improve upon the mention-only baseline by substituting proforms, e.g. pronouns such as *"it"*, and mentions without descriptive content, e.g. phrases such as *"the one you mentioned"*, with the most recent mention that does not belong to either category. This is expected to be a relatively strong baseline when mentions are specific and anaphora have mostly local antecedents.

3.2.2 Coreference Resolution

We opt for an off-the-shelf[2] span-based coreference resolution model (**coref**) originally presented in Lee et al. (2018), but that has since been updated to use SpanBERT (Joshi et al., 2020) instead of the original GloVe embeddings (Pennington et al., 2014). For each mention, we use the model to resolve its coreference links and aggregate all coreferential information in its cluster based on the given context window.

[2] https://github.com/allenai/allennlp-models/tree/main/allennlp_models/coref

We experiment with two different representations of the referent descriptions from this model, those being (1) a concatenation of all of the mention's coreferences and (2) an ordered *set-of-words* representation that contains only the unique lexical items in the cluster. To offset that this model was not specifically trained to handle coreference in conversation, we provide it with the contents of the span of the mention when it does not manage to detect the mention itself and, consequently, not connect it to any of its coreferences. For partial matches, in addition to adding all tokens from the cluster associated with the match, we also add the missing tokens from the span to the description.

4 Experiments

4.1 Data

We use the dialogues from the collaborative image ranking task **A Game Of Sorts** (AGOS, Willemsen et al., 2022) for our experiments. In AGOS, two players are asked to rank a set of images based on a given sorting criterion. They see the same set of images, but the position of the images on the screen is randomized for each player. Through a largely unrestricted conversation, and without being able to see the perspective of the other player, the players need to agree on how to rank the images given the sorting criterion. Sorting criteria are embedded in scenarios that are intended to create a discussion, leading to mixed-initiative interactions with both parties contributing to the discourse. Each interaction takes place over four rounds with the same set of nine images, effectively guaranteeing repeated references. The image sets used for the game cover five different image categories. Each set contains nine images with each image representing an entity from one of these categories as its main subject. Willemsen et al. (2022) collected three interactions per image set for a total of 15 dialogues.

Ground Truth Our formulation of the visually-grounded reference resolution problem requires span-based annotations of mentions aligned with the image(s) they denote. These annotations are the basis of our ground truth references for both training and evaluation. We follow Willemsen et al. (2022) regarding the marking of mentions in AGOS, in that we only annotate those that are either singletons or are part of an identity relation with other mentions that have an exophoric referent that is part of the visual context, i.e. regardless of form, any referring expression that is meant to denote one or more of the images. During the game, players were asked to provide self-annotations: for each message they sent they were asked to indicate which image(s), if any, they were referring to. We use these self-annotations, post-edited where necessary, to manually mark the spans of mentions that can be grounded in the visual context.

We create three different representations of the ground truth referent descriptions. Two are automatically extracted from the marked mentions and are similar in structure to the labels of the **coref** baseline, i.e. (1) an incremental concatenation of the reference chain and (2) an incremental ordered set of words consisting of the unique lexical items in the cluster. The third are manually constructed labels that summarize reference chains as definite descriptions. For each representation, the context window dictates which references are considered for the label.

4.2 Model Specifications

For pointers to implementations, we refer the reader to our repository[3].

4.2.1 LLMs

We fine-tune two LLMs, GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), for conversational referent description generation. For hyperparameters, see our Supplementary Material.

GPT-2 We fine-tune the 1.5 billion parameter GPT-2 model.

GPT-3 We fine-tune the 175 billion parameter `davinci` base model using the OpenAI API.

4.2.2 VLMs

We evaluate the zero-shot text-image retrieval performance of several pretrained VLMs for our task, those being CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), BLIP (Li et al., 2022), and BLIP-2 (Li et al., 2023).

CLIP We evaluate two variants of CLIP, CLIP ViT-B/16 and CLIP ViT-L/14.

ALIGN We use the COYO-ALIGN implementation trained from scratch on COYO-700M.

BLIP We use the BLIP base model.

BLIP-2 We use the BLIP-2 model that was fine-tuned on the Karpathy and Fei-Fei (2015) training set split of MS COCO (Lin et al., 2014).

[3] Available upon publication.

4.3 Evaluation

We perform (nested) five-fold cross-validation by partitioning the AGOS dataset along the five image sets. To avoid leakage, for each run we use the three dialogues from one image set as the held out test set and train on the twelve dialogues from the four other image sets. To evaluate how dialogue history affects results, we report performance of the different methods for two context windows, **3** and **7**. In addition, we examine whether increasing the size of the context window further would, in principle, lead to greater returns, by assessing ground-truth performance for windows of **13** and the **full** dialogue context. Finally, we conduct an error analysis of the generated descriptions.

Note that because our model does not have access to game state information with respect to the visual context, we make a simplifying assumption with regard to the images and reduce the candidate set as the game progresses. A successfully ranked image is no longer considered part of the visual context for that round. Although this does mean that the models will not be able to identify the referent for references to ranked images, as they will not be part of the candidate set, such references are an extremely rare occurrence, as players must discuss the unranked images to progress with the task. For the sake of completeness, we will also report results for the unchanged candidate set.

4.3.1 Metrics

We measure task success for visually-grounded reference resolution in terms of text-image retrieval performance. In addition, we estimate the quality of the referent descriptions by comparing them to the manually constructed ground truth labels using text similarity metrics.

Text-Image Retrieval We estimate the image retrieval performance based on accuracy $[0, 1]$, mean reciprocal rank (MRR) $[0, 1]$, and normalized discounted cumulative gain (NDCG) $[0, 1]$. We limit our evaluation to single-image referents. Accuracy is top-1 accuracy.

For our random lower bound, we can calculate the expected values for accuracy and MRR. For top-1 accuracy we take 1 over the size of the set of candidate images per item, averaging over all items. For MRR we take 1 over the size of the set of candidate images, divided by two per item, averaging over all items. Calculating an expected value for NDCG of a random model is intractable due to its dependence on relevancy scores.

Text Generation We evaluate the output from the CRDGs by comparing the generated descriptions to the manually constructed ground truth labels using metrics to quantify similarity. We use the Jaccard index $[0, 1]$ to assess vocabulary overlap. We use BLEU $[0, 1]$ (Papineni et al., 2002) to assess similarity based on n-gram overlap (unigrams to four-grams). We use the longest common subsequence variant of ROUGE $[0, 1]$ (Lin, 2004), i.e. ROUGE-L, as a further indication of the preservation of word order. In addition, we opt for an embedding-based metric as a proxy for semantic equivalence between the predicted and reference sequences. For this purpose, we compute cosine similarity $[0, 1]$ between text embeddings.

4.3.2 Human

We conduct two different human subject experiments to assess human performance for this task. We provide additional details about the experimental setup in the Supplementary Material.

Independent We conduct an experiment aimed at comparing VLM and human performance on the task where every trial is independent. Participants are given a referent description and are asked to select from a set of candidate images the image they believe is best described by the label. The images and labels are presented to the participants independent of the dialogue. Note that we evaluate with the reduced candidate set. The referent descriptions are the manually constructed ground truth labels based on the **full** dialogue context. To collect data for all labels, ensuring independence of observations, we recruited 322 participants via crowdsourcing. The crowdworkers were financially compensated for their contributions.

Holistic We conduct an experiment in which mentions are shown to participants within the context of the dialogue. For each mention, the participants are presented with the dialogue leading up to and including the message which contains the reference. The start and end of the span of the mention that the participant is asked to resolve are visually indicated. For each marked mention, the participant is asked to select which image or images are referenced. As they progress with the task, participants will have access to increasingly more of the dialogue history. For each mention the participants are presented with all images, but with a visual indication of their status, i.e. for each image whether the players had managed to successfully rank it

	Accuracy		MRR		NDCG	
	3	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.43	.43	-	-
Mention	.59	.59	.73	.73	.79	.79
Substitution	.68	.68	.80	.80	.85	.85
coref, chain	.65	.66	.78	.79	.83	.84
coref, set	.66	.66	.78	.79	.84	.84
GT, chain	.73	.74	.83	.85	.87	.88
GT, set	.73	.75	.84	.85	.87	.89
GT, manual	.72	.74	.83	.84	.87	.88
GPT-2	.64	.60	.74	.76	.83	.80
GPT-3	.69	.71	.81	.82	.86	.86

Table 1: Cross-validated image retrieval performance averaged over five folds for single-image referents. *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth.

	GPT-2		GPT-3	
	3	**7**	**3**	**7**
BLEU	.55	.47	.75	.70
ROUGE-L	.71	.65	.86	.83
Jaccard	.44	.35	.70	.63
Cosine	.88	.85	.96	.95

Table 2: Text generation metrics evaluation results averaged over five folds for single-image referents. *Note*. Scores are rounded to the nearest hundredth.

at that point in the interaction. We recruited 23 participants via crowdsourcing. For each of the 15 AGOS dialogues we collected data from two different participants. Each participant was allowed to provide data for at most one dialogue per image set. The crowdworkers were financially compensated for their contributions.

5 Results

5.1 Text-Image Retrieval

Table 1 shows, for context windows **3** and **7**, the zero-shot text-image retrieval performance results for the VLM that averaged best performance over the five folds, namely BLIP-2. For the text-image retrieval accuracy achieved by the other VLMs, performance on the unreduced candidate set, and accuracy per fold for BLIP-2, see Appendix C.

As can be seen from the results presented in Table 1, we achieve best performance with a fine-tuned GPT-3 as the CRDG and BLIP-2 for zero-shot text-image retrieval. In addition to outperforming the baselines, we find that GPT-3 is a more performant discourse processor for this task than GPT-2. This result is consistent between the VLMs.

Results generally show a slight increase in performance when increasing the context window from **3** to **7**. Performance on the ground truth reference descriptions for context windows **13** and the **full** dialogue shows this trend persists, with BLIP-2 achieving approximately 75% and 82% accuracy, respectively. A plot of the performance for the four context windows is shown in Figure 6 in Appendix C. This result suggests that the size of the context window may have a significant impact on performance, with a 10% increase in accuracy from **3** to **full**. Furthermore, the VLMs do not seem overly sensitive to the composition of the referent descriptions, as performance is largely comparable between the automatically generated and the manually constructed ground truth labels.

We find that BLIP-2 is on par with human text-image retrieval performance in terms of top-1 accuracy for the manually constructed ground truth referent descriptions based on the full dialogue history for single-image referents, as our human participants averaged roughly 80% accuracy in the independent setup. However, when we compare these results with the single-image referent text-image retrieval performance in the holistic setup, we see that the upper bound for this task when references are resolved within the combined linguistic and extralinguistic dialogue context is likely considerably higher as our human participants averaged approximately 91% accuracy (average of best performance per dialogue is roughly 93%).

5.2 Text Generation

Table 2 shows the text generation metric results averaged over the five folds, providing an indication of the extent to which the fine-tuned LLMs managed to generate referent descriptions that approximate the manually constructed ground truth labels. We observe that an increase in context window size results in a decrease in scores, which is consistent across metrics. Interestingly, we did not find such a decrease with respect to text-image retrieval performance. We do again find GPT-3 to be more performant than GPT-2, here in terms of approximating the ground truth.

5.3 Error Analysis

Examining the output from the fine-tuned GPT-3 model, we observe a number of recurring errors.

The most notable errors are those where the model fails to link a mention to (all of) its coreferences that are present in the dialogue segment, or links mentions that denote different referents. For

example, for one mention the ground truth label is *"the sheep dog"*, but the generated label was *"the sheep dog with a leash"*; the model incorrectly attributed the prepositional phrase to the mention as it was actually a descriptor for a different referent. Related, since the CRDGs function at the message level, a mention can have both anaphoric and cataphoric coreferences when there are multiple mentions of the same referent in an utterance An example of such an utterance is *"Good question. I think the angry one also looks a little wild. So that could be an option as well. I mean the one with white nose and forehead"*, where *"the angry one"*, *"that"*, and *"the one with white nose and forehead"* are all mentions of the same referent with the same ground truth label *"the angry dog with a white nose and forehead"*. The model generates this correctly for the latter two, but not the former one for which only *"the angry dog"* was generated, meaning it correctly substituted the proform but did not link the mention with its cataphoric coreferences.

Finally, some generated referent descriptions differ from the ground truth in terms of lexical choice or syntax, but not in terms of information content. This negatively affects scores of text generation metrics based on overlapping content in particular, but these are otherwise not meaningful errors as there are multiple ways to construct semantically similar descriptions, e.g., *"the big dog which looks scary"* versus *"the big scary-looking dog"*.

6 Discussion

We have presented an approach to visually-grounded reference resolution that frames the discourse processing side of the task as a causal language modeling problem. By fine-tuning an LLM to generate referent descriptions for marked mentions in dialogue segments from the collaborative image ranking task A Game Of Sorts (Willemsen et al., 2022), we demonstrate the possibility of treating referent identification as a zero-shot text-image retrieval problem by using pretrained VLMs for the grounding of the generated labels. As we have not in any way indicated coreferential relations in the fine-tuning training data, our results imply that certain pretrained LLMs, here GPT-3, may learn to resolve coreferences implicitly without the need for explicit supervision for this fundamental subtask.

In this work, we have treated the processing of the discourse as entirely disjoint of the visual modality. As such, it has inherent limitations. The mentions we find in the dialogues have not been produced void of the extralinguistic context. The dialogue participants could rely on co-observed visual stimuli to help resolve otherwise ambiguous language use. From linguistic context alone, some ambiguities, such as prepositional phrase attachment, may be impossible to resolve. It is, therefore, noteworthy that the downstream zero-shot text-image retrieval performance using the generated descriptions from our unimodal approach far exceeds chance level accuracy, with the potential for results to improve further given access to the full dialogue history, as we found that the ground truth labels based on larger context windows achieve greater text-image retrieval performance. However, the results from our holistic human evaluation support the notion that a multimodal approach should ultimately prove even more effective.

We found that a decrease in text generation metric scores did not necessarily indicate a similar decrease in text-image retrieval performance, suggesting that the generated descriptions captured sufficiently discriminative information about the referents and achieved similar grounding accuracy despite not approximating the ground truth labels to the same extent. It is also important to note that mentions may not have a single, canonical ground truth referent description due to lexical and syntactic variations between referring attempts.

Despite the relatively small size of the dataset collected by Willemsen et al. (2022), we were still able to fine-tune GPT-3 to perform the task with greater accuracy than the baselines, which speaks to the sample efficiency of pretrained LLMs. In comparison, we find that the much smaller GPT-2 is prone to intrusions from the fine-tuning training data and more often fails to resolve the coreferences correctly. Although the complexity of the discourse warrants the use of more powerful models, it is, nevertheless, likely that any LLM used for the task would benefit from a larger fine-tuning dataset. Related, benchmarking performance on other visually-grounded dialogue tasks would provide insights into the generalizability of the method.

In addition to pursuing a multimodal approach, finer-grained grounding, and evaluating our method on other datasets, possible avenues for future work include expanding the annotations to include coreferential relations other than identity relations, addressing multi-image referents, and unifying the method with a mention proposal system.

References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for Visual Dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197, Online. Association for Computational Linguistics.

Santanu Bhattacharjee, Rejwanul Haque, Gideon Maillette de Buy Wenniger, and Andy Way. 2020. Investigating Query Expansion and Coreference Resolution in Question Answering on BERT. In *Natural Language Processing and Information Systems*, pages 47–59, Cham. Springer International Publishing.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. PaLM: Scaling Language Modeling with Pathways. _eprint: 2204.02311.

Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual Referring Expression Recognition: What Do Systems Actually Learn? 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 781–787, New Orleans, Louisiana. Association for Computational Linguistics.

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

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1080–1089.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual Object Discovery through Multi-modal Dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meet Up! A Corpus of Joint Activity Dialogues in a Visual Environment. In *Proceedings of the 23rd Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, London, United Kingdom. SEMDIAL.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Andrej Karpathy and Li Fei-Fei. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Casey Kennington, Livia Dia, and David Schlangen. 2015. A Discriminative Model for Perceptually-Grounded Incremental Reference Resolution. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 195–205, London, UK. Association for Computational Linguistics.

Casey Kennington and David Schlangen. 2017. A simple generative model of incremental reference resolution for situated dialogue. *Computer Speech & Language*, 41:43–67.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-Order Coreference Resolution with Coarse-to-Fine 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 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

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. _eprint: 2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. 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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. 2017. Recurrent Multimodal Interaction for Referring Image Segmentation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1280–1289.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jun Quan, Deyi Xiong, Bonnie Webber, and Changjian Hu. 2019. GECOR: An End-to-End Generative Ellipsis and Co-reference Resolution Model for Task-Oriented Dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4547–4557, Hong Kong, China. Association for Computational Linguistics.

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, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and others. 2018. Improving language understanding by generative pre-training. Publisher: OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

David Schlangen. 2019. Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings. *CoRR*, abs/1908.11279. ArXiv: 1908.11279.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1213–1223, Berlin, Germany. Association for Computational Linguistics.

Todd Shore, Theofronia Androulakaki, and Gabriel Skantze. 2018. KTH Tangrams: A Dataset for Research on Alignment and Conceptual Pacts in Task-Oriented Dialogue. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Todd Shore and Gabriel Skantze. 2018. Using Lexical Alignment and Referring Ability to Address Data Sparsity in Situated Dialog Reference Resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2288–2297, Brussels, Belgium. Association for Computational Linguistics.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. _eprint: 2302.13971.

Takuma Udagawa and Akiko Aizawa. 2019. A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press. Event-place: Honolulu, Hawaii, USA.

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.

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.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling Context in Referring Expressions. In *Computer Vision – ECCV 2016*, pages 69–85, Cham. Springer International Publishing.

Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. 2018. Grounding Referring Expressions in Images by Variational Context. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4158–4166.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. _eprint: 2205.01068.

Appendices

A Dialogue Excerpt

the blackberry with the keyboard looks older than others

yes and I think either the sliver one or the black one is okay

the two are very similar

lets rank the nokia No3 and the blackberry No4

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

I am talking about the silver Nokia and the silver blackberry

Figure 3: Excerpt of an AGOS dialogue with references to single-image referents underlined; the color indicates the referent. *Note*. The two images that have been ranked successfully at this point in the interaction have a faded appearance.

B Model Input

```
A:    Which one do you think is the smallest?
B:    the black ear beagle!                    -> the beagle dog with black ears
B:    looking to the left                      -> the beagle dog with black ears looking to the left
A:    The one ears longer than the head?       -> the beagle dog with black ears longer than its head looking to the left
B:    yes
```

Figure 4: Excerpt of an AGOS dialogue with messages paired with manually constructed ground truth referent descriptions. Mentions are in purple and made bold for illustrative purposes. Ground truth labels for the referent denoted by the mention in green.

```
                    "Speaker" token + task instructions   <N> 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?
                                Speaker token + utterance <A> both of them seem quite hard to use. Yeah let's choose bigger one first
                                Speaker token + utterance <B> I heard Nokia is pretty solid
                                Speaker token + utterance <A> yeah I was thinking the same
      Speaker token + utterance with a marked mention   <B> Maybe >>the one with rubber<< ? Easier to grab
                                      Inference token   ->
Ground truth referent description for the marked mention the phone with rubber END
```

Figure 5: Sample input to LLM, deconstructed for demonstration purposes (the sample is otherwise a flat sequence of tokens). Left (text in purple): explanation of input; right (text in black): input. *Note*. The ground truth is only available to the model during training, not during inference.

C Additional VLM Results

	CLIP-B		CLIP-L		ALIGN		BLIP	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.11	.11	.11	.11	.11	.11	.11	.11
Mention	.36	.36	.44	.44	.44	.44	.40	.40
Substitution	.42	.42	.51	.51	.52	.52	.50	.50
coref, chain	.41	.42	.49	.49	.47	.46	.47	.46
coref, set	.42	.41	.48	.48	.49	.48	.47	.47
GT, chain	.45	.47	.54	.56	.53	.53	.52	.54
GT, set	.46	.48	.54	.56	.54	.54	.53	.55
GT, manual	.47	.48	.53	.55	.58	.59	.55	.57
GPT-2	.41	.38	.46	.43	.49	.44	.47	.43
GPT-3	.44	.45	.52	.52	.54	.55	.52	.52

Table 3: Cross-validated image retrieval accuracy averaged over five folds for single-image referents (candidate set not reduced). *Note*. Scores are rounded to the nearest hundredth. GT = ground truth; CLIP-B = CLIP ViT-B/16; CLIP-L = CLIP ViT-L/14.

	CLIP-B		CLIP-L		ALIGN		BLIP	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.22	.22	.22	.22	.22	.22
Mention	.49	.49	.55	.55	.56	.56	.54	.54
Substitution	.55	.55	.62	.62	.64	.64	.64	.64
coref, chain	.54	.54	.61	.61	.60	.60	.61	.61
coref, set	.54	.53	.61	.60	.61	.61	.61	.61
GT, chain	.58	.59	.66	.67	.66	.67	.66	.68
GT, set	.58	.60	.66	.68	.67	.67	.66	.69
GT, manual	.59	.60	.64	.66	.69	.70	.69	.70
GPT-2	.53	.49	.58	.54	.61	.58	.60	.58
GPT-3	.57	.58	.63	.63	.66	.66	.67	.68

Table 4: Cross-validated image retrieval accuracy averaged over five folds for single-image referents (candidate set reduced). *Note*. Scores are rounded to the nearest hundredth. GT = ground truth; CLIP-B = CLIP ViT-B/16; CLIP-L = CLIP ViT-L/14.

	Cars		Dogs		Paintings		Pastries		Phones	
	3	**7**	**3**	**7**	**3**	**7**	**3**	**7**	**3**	**7**
Random	.22	.22	.22	.22	.22	.22	.22	.22	.22	.22
Mention	.52	.52	.62	.62	.60	.60	.61	.61	.58	.58
Substitution	.63	.63	.70	.70	.70	.70	.68	.68	.67	.67
coref, chain	.59	.60	.69	.69	.66	.67	.67	.68	.63	.63
coref, set	.60	.57	.68	.68	.69	.68	.69	.70	.62	.62
GT, chain	.66	.66	.76	.78	.72	.74	.75	.78	.71	.69
GT, set	.66	.65	.74	.77	.73	.78	.76	.80	.73	.73
GT, manual	.64	.63	.75	.78	.77	.80	.70	.72	.74	.74
GPT-2	.62	.62	.67	.62	.67	.62	.63	.61	.57	.50
GPT-3	.63	.63	.75	.78	.70	.70	.68	.72	.70	.69

Table 5: Cross-validated image retrieval accuracy per fold for single-image referents (candidate set not reduced). *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth; CLIP-B = CLIP ViT-B/16; CLIP-L = CLIP ViT-L/14.

	Accuracy		MRR		NDCG	
	3	**7**	**3**	**7**	**3**	**7**
Random	.11	.11	.22	.22	-	-
Mention	.47	.47	.63	.63	.72	.72
Substitution	.55	.55	.71	.71	.78	.78
coref, chain	.53	.51	.69	.68	.76	.76
coref, set	.53	.51	.69	.68	.77	.76
GT, chain	.60	.61	.75	.76	.81	.82
GT, set	.60	.62	.75	.77	.81	.83
GT, manual	.63	.64	.76	.78	.82	.83
GPT-2	.54	.48	.69	.65	.77	.73
GPT-3	.60	.60	.74	.74	.80	.81

Table 6: Cross-validated image retrieval performance averaged over five folds for single-image referents (candidate set not reduced). *Note*. Scores shown are of VLM that averaged best performance (BLIP-2). Scores are rounded to the nearest hundredth. GT = ground truth.

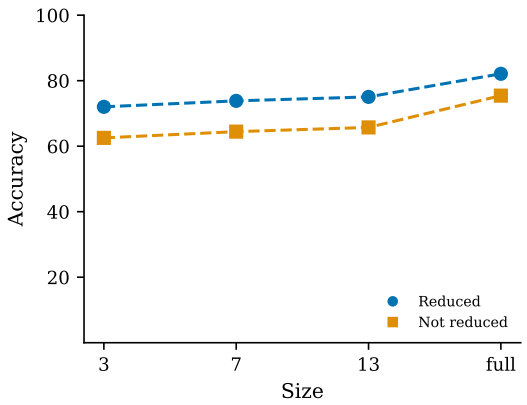

Figure 6: Text-image retrieval accuracy as a function of the size of the context window. Results are shown for BLIP-2 on the manually constructed ground truth referent descriptions based on their respective windows. We show results for both the reduced candidate set and the not reduced candidate set.