UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Dealing with Semantic Underspecification in Multimodal NLP

Pezzelle, S.

DOI
10.18653/v1/2023.acl-long.675
Publication date
2023
Document Version
Final published version
Published in
The 61st Conference of the Association for Computational Linguistics

License
CC BY

Link to publication

Citation for published version (APA):
Pezzelle, S. (2023). Dealing with Semantic Underspecification in Multimodal NLP. In A. Rogers, J. Boyd-Graper, & N. Okazaki (Eds.), *The 61st Conference of the Association for Computational Linguistics: ACL 2023 : Proceedings of the Conference : July 9-14, 2023* (Vol. 1, pp. 12098-12112). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.675

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Oct 2026

Dealing with Semantic Underspecification in Multimodal NLP

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Intelligent systems that aim at mastering language as humans do must deal with its *semantic underspecification*, namely, the possibility for a linguistic signal to convey only part of the information needed for communication to succeed. Consider the usages of the pronoun *they*, which can leave the gender and number of its referent(s) underspecified. Semantic underspecification is not a bug but a crucial language feature that boosts its storage and processing efficiency. Indeed, human speakers can quickly and effortlessly integrate semantically-underspecified linguistic signals with a wide range of non-linguistic information, e.g., the multimodal context, social or cultural conventions, and shared knowledge. Standard NLP models have, in principle, no or limited access to such extra information, while multimodal systems grounding language into other modalities, such as vision, are naturally equipped to account for this phenomenon. However, we show that they struggle with it, which could negatively affect their performance and lead to harmful consequences when used for applications. In this position paper, we argue that our community should be aware of semantic underspecification if it aims to develop language technology that can successfully interact with human users. We discuss some applications where mastering it is crucial and outline a few directions toward achieving this goal.

1 Introduction

They put the flowers there. Speakers of a language hear sentences like this every day and have no trouble understanding what they mean—and what message they convey. This is because, in a normal state of affairs, they can count on a wide range of information from the surrounding context, personal knowledge and experience, social or cultural conventions, and so on. Upon hearing this sentence, for example, they would know that flowers go into vases, look in the direction where their interlocutor nodded their chin, see a vase with tulips on the windowsill, and *infer* that this is where someone put the flowers. Every time listeners need to count on extra, non-linguistic information to understand a linguistic signal, like in this example, it is because the language used is semantically *underspecified* (Ferreira, 2008; Frisson, 2009; Harris, 2020b). In the example above, the locative adverb *there* leaves underspecified a location—where the flowers were put—which would instead be explicitly provided in the semantically more specified sentence *They put the flowers in the light blue vase on the windowsill at the end of the hallway*. According to linguists, indeed, adverbs of place (*here*, *there*) are typical examples of semantically underspecified words, as well as adverbs of time (*now*, *today*), demonstratives (*this*, *that*), quantifiers (*few*, *many*), tensed expressions, and some usages of personal pronouns (Lappin, 2000; Harris, 2020b).

The reason why semantic underspecification is so widespread has to do with language efficiency, which is a trade-off between informativeness and conciseness (Zipf, 1949; Goldberg and Ferreira, 2022). Underspecified words can be used in many communicative occasions with varying meanings and intentions (Harris, 2020b), which prevents speakers from fully articulating every nuance of meaning every time they talk (Piantadosi et al., 2012). Indeed, planning and producing utterances—but also speech (see Levinson, 2000)—is cognitively expensive (Trott and Bergen, 2022). The use of underspecified language, at a first sight, seems to go against the view that language is a cooperative system (Grice, 1975; Tomasello, 2005) and can indeed explain cases where communication appears to be *egocentric* rather than cooperative (Keysar, 2007). However, a wealth of studies has shown that humans are extremely good at making inferences (Grice, 1969; Sedivy et al., 1999) and that this ability is cognitively cheaper than speaker articulation, which is rather demanding and time-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 12098–12112
July 9-14, 2023 ©2023 Association for Computational Linguistics

consuming (Levinson, 2000). Upon hearing a semantically underspecified sentence, human speakers can quickly and effortlessly integrate linguistic and non-linguistic information (Harris, 2020b). In this light, Levinson (2000) proposed that semantic underspecification gives rise to processing efficiency besides boosting storage efficiency.

Semantic underspecification allows our limited repertoire of symbols to be used in many contexts and with different intentions without compromising its communicative effectiveness. For example, we can use the pronoun *they* to omit a person's gender or refer to a group of friends; the locative *here* to refer to a free table at a café or the institution you work for. Semantic underspecification is not a bug but a crucial feature of language that is ubiquitous in human communication (Harris, 2020a). As such, any intelligent system that aims at mastering language as humans do must be able to properly deal with it. This is particularly the case for models of natural language understanding and generation that have access to non-linguistic information (Bender and Koller, 2020; Bisk et al., 2020), e.g., models integrating language and vision that account for the multimodality of language (Harnad, 1990). These models must be able to understand and generate sentences like *They put the flowers there*, provided that a relevant visual context is present and there is a clear communicative goal. This is a mandatory requirement if we want to use these systems to model real communicative scenarios or embed them in applications that interact with human users.

In this position paper, we argue that semantic underspecification should be high on the NLP community agenda, particularly within approaches combining language and vision. We report that SotA multimodal NLP models struggle with it, and advocate a comprehensive, thorough investigation of the phenomenon along several research directions and concrete steps. Mastering semantic underspecification is a long-term goal that implies shifting the paradigm to a scenario where models use language as humans do, that is, *with a communicative goal*. In line with what was argued elsewhere (Bisk et al., 2020; Giulianelli, 2022; Fried et al., 2022), we believe the time is ripe for such a change.

2 How Do Multimodal Models Deal with Semantic Underspecification?

The field of multimodal or visually grounded NLP is currently dominated by pre-trained multimodal Transformers. Since their introduction, models like CLIP (Radford et al., 2021), LXMERT (Tan and Bansal, 2019), VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), VL-BERT (Su et al., 2019), UniT (Hu and Singh, 2021), VILLA (Gan et al., 2020), UNITER (Chen et al., 2020), VinVL (Zhang et al., 2021), ViLT (Kim et al., 2021), and mPLUG (Li et al., 2022), *inter alia*, have rapidly become the new state-of-the-art in virtually every language and vision task. Among other tasks, these models achieve unprecedented performance in describing an image in natural language (Lin et al., 2014), finding the best image for a given language query (Plummer et al., 2015), answering fine-grained questions about the content of an image (Antol et al., 2015; Krishna et al., 2017; Hudson and Manning, 2019), reasoning over objects and object relations (Johnson et al., 2017; Suhr et al., 2019), and entertaining a visually-grounded dialogue by asking and answering questions (De Vries et al., 2017; Das et al., 2017).

These models differ from each other in several dimensions. For example, they either concatenate and jointly process the visual and textual embeddings (*single-stream* models), or process the two modalities by means of separate encoders with an optional cross-modal fusion (*dual-stream* models); or, they use visual features extracted with either CNN-based (e.g., region features from Faster R-CNN; Ren et al., 2015) or Transformer-based (e.g., image features from Vision Transformer, ViT; Dosovitskiy et al., 2020) image encoders. However, they share both the same underlying architecture, which is based on Transformers, and training regime, which leverages a massive amount of multimodal data and a few common learning objectives. One of the most popular learning objectives is Image-Text Matching (ITM), which maximizes the similarity between an image and a language fragment that is well *aligned* with it. As a result of this training regime, these models are impressively good at judging whether a sentence is a good/bad (true/false) description of the content of an image. This is particularly the case for CLIP, which is optimized for the task and can almost perfectly spot word-level inconsistencies between an image and a sentence, as the ones included in the FOIL dataset by Shekhar et al. (2017) (results reported in Parcalabescu et al., 2022).

Given this impressive performance, it is reasonable to expect that these models are robust to se-

mantically underspecified language. Describing an image, asking a question, or entertaining a conversation about it are all communicative scenarios that admit a varying degree of semantic underspecification. For example, the question *What are they doing?* referred to a visual context with people playing an unusual sport is perfectly acceptable—and indeed likely to be asked; or, the sentence *A person is typing on their laptop* to describe an office environment is not only a very good description of that context but perhaps even a desirable one. Therefore, mastering semantically underspecified language is a requisite for any multimodal NLP model which aims at both genuinely solving these tasks and being used for user-facing applications.

2.1 Proofs of Concept

To scratch the surface of the problem, we carry out two Proofs of Concept (hence, PoCs) using image descriptions and the CLIP model. When talking about a visual context, speakers of a language can convey the same message with varying levels of semantic specification. For example, they can describe someone waiting for the bus by referring to them as *an elderly lady*, *a woman*, *a person*, or *they*. Similarly, they can mention a location, i.e., *the bus stop*, or use the locatives *here* or *there*; an object, i.e., *the bus*, or use the demonstratives *this* or *that*; and so on. This is possible because the visual context provides enough information for the addressee to understand the message, even when it is extremely semantically underspecified.

Almost by definition, standard image descriptions as those in COCO (Lin et al., 2014) are semantically *overspecified*. Indeed, they are meant to serve as a natural language 'translation' of the content of an image, to make it available to someone who does not have access to the image (for a discussion on this point, see, e.g., Kreiss et al., 2022b). As such, these descriptions fully specify a wide range of semantic aspects that would be reasonably left underspecified if the visual context was available to both interlocutors. As mentioned above, CLIP is extremely good at assessing whether a description is good for an image. As such, it is reasonable to expect that the model should not be affected by the degree of semantic specification of the description, provided that it is valid for the image. To illustrate, a model should similarly score the descriptions *A woman waiting for the bus* and *A person waiting for the bus* in relation to the visual context described above. Moreover, a semantically valid underspecified description must always be better than an unrelated, overspecified description.

In the two PoCs below, we explore these two hypotheses. Note that we do so for illustrative purposes, highlighting general trends that can be useful for further, more thorough research. Moreover, it is worth stressing that, while we employ CLIP due to its effectiveness and accessibility, the point we make is more general in scope than focused on this specific model. The point is that *models should not be affected by semantic underspecification when assessing the validity or applicability of an image description.* Concretely, we use 100 images and corresponding descriptions (495 in total) from the 2014 train partition of COCO. Data and code available at: `https://github.com/sandropezzelle/sunglass`

Are Underspecified Descriptions as Good as Overspecified Ones? In this PoC, we are interested to check whether CLIP is robust to semantic underspecification. The expectation is that the model should assign the same or a similar alignment score to image descriptions with a varying level of semantic specification, provided that these descriptions are semantically correct for the image.

We compute CLIPScore for each of the 495 ⟨image, description⟩ pairs in our sample and select the 100 with the highest score. We refer to these 100 descriptions as `Original`. We then create up to 6 underspecified versions of each description in `Original` by manually perturbing their text to account for various underspecification phenomena. Such an annotation task was performed by a single annotator, the author of this paper, with a background in formal and computational linguistics. Perturbations are carried out only where possible (thus, not all descriptions have all 6 versions), without altering the grammatical structure of the sentence. The semantic underspecification phenomena we consider are illustrated in the example in Figure 1 and described below:

- `Quantity:` We replace numbers (e.g., *two*) and quantity expressions (e.g., *a couple*) with the quantifier *some*
- `Gender:` We replace gender-marked (e.g., *woman*) and age-marked (e.g., *children*) nouns with the hypernyms *person* or *people*
- `Gender+Number:` We replace any NPs in subject position, either singular or plural, with the

TYPE	DESCRIPTION	SCORE
`Original`	The woman is standing above the two packed suitcases.	0.8565
`Quantity`	The woman is standing above **some** packed suitcases.	0.8275
`Gender`	The **person** is standing above the two packed suitcases.	0.7608
`Gender+Number`	**They** are standing above the two packed suitcases.	0.7435
`Location`	The woman is standing **here**.	0.5537
`Object`	The woman is standing above **this**.	0.4931
`Full`	**They are doing something here.**	0.4646

Figure 1: For one image, `COCO_train2014_000000394151.jpg`, we report one original description from COCO and the six corresponding semantically underspecified descriptions, along with their CLIPScore.

pronoun *they*, and harmonize verb agreement

- `Location`: We replace PPs introduced by a preposition of place (e.g., *at*) with the locatives *here* or *there*
- `Object`: We replace NPs, typically in object position, with the demonstratives *this* or *that*
- `Full`: We replace the entire sentence with the fully underspecified one *They are doing something here.*

We compute CLIPScore for each underspecified description. In Figure 2, we report the distribution of these scores against each phenomenon. As can be seen, underspecified descriptions achieve (much) lower scores compared to `Original` ones. For example, a perturbation as harmless as replacing the subject with the pronoun *they* leads to a ∼16-point average decrease in CLIPScore, while the gap increases to ∼40 points when considering `Original` against the fully underspecified description *They are doing something here.* All the scores for one specific example are reported in Figure 1.

These observations are surprising and go against our expectations that underspecified descriptions, if semantically valid, should be considered as good as overspecified ones. Indeed, *why should a sentence containing a quantifier, a pronoun, or a locative be considered a poor description of a visual context?* One possible explanation is that models like CLIP are sensitive to the amount of *detail* provided by an image description. More specifically, the more words there are in the sentence with a clear and unique visual referent, the more the description is deemed 'aligned' to an image. Using the terminology introduced by Kasai et al. (2022) to evaluate image captioning metrics, the model would be good at capturing an image description's *recall*, i.e., the extent to which the salient objects in an image are covered in it; on the other hand, it would poorly

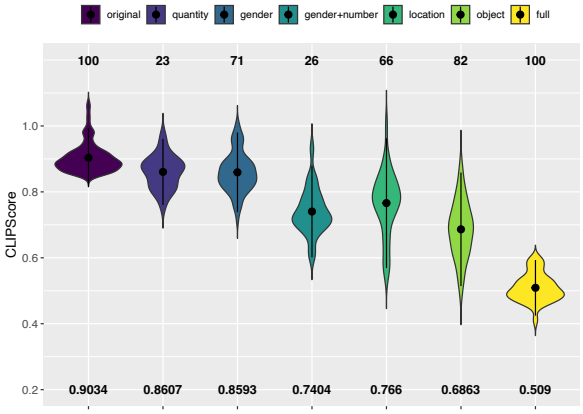

Figure 2: Violin plot reporting the distribution of CLIPScore values per image description type. The dot and bars in each violin stand for the mean and standard deviation, respectively. The mean of each violin is also given at the bottom of the plot, while the integer at the top reports the sample size. Best viewed in color.

capture a description's *precision*, i.e., the degree to which it is precise or valid for a given image.[1] If this was the case, models like CLIP would end up always considering underspecified descriptions as worse than overspecified ones, which naturally raises questions about their robustness and applicability to a possibly wide range of scenarios.

Are Underspecified Descriptions Better than Unrelated Ones? Even if CLIP was sensitive to the amount of detail provided by an image description (the more, the better), a valid underspecified description should always be deemed more aligned than an unrelated, overspecified one. That is, even a highly underspecified sentence like *They are doing something here*—if semantically valid for the image, which is the case in our small sample—should always be preferred over a description that is fully unrelated to the image. To test this hypothesis, we experiment with this `Full` description,

[1]We thank the anonymous reviewer who referred us to this work.

`Full` They are doing something here.

`r1` A woman in a white dress is sitting with her cell phone.
`r2` A girl with long brown hair with streaks of red lays on a bed and looks at an open laptop computer.
`r3` A lady holding a bottle of ketchup and a dog in a hot dog bun costume.
`r4` An infant sits next to a stuffed teddy bear toy.
`r5` Woman sitting on a bench holding a hotdog in her hand
`r6` Two small children playing with their refrigerator magnets.

Figure 3: For `COCO_train2014_000000205931.jpg`, our fully underspecified image description achieves a lower CLIPScore than six randomly picked overspecified descriptions, although these are clearly wrong.

and, for each image, we test it against 10 randomly sampled `Original` descriptions of other images. Surprisingly, for 82 images out of 100, at least one random caption achieves a higher CLIPScore than `Full`. While the actual numbers may depend on the number and the type of random descriptions being sampled, some qualitative observations are helpful to highlight the behavior of the model. Consider, as an example, the case reported in Figure 3, where 6 out of 10 unrelated descriptions are deemed better than our fully underspecified one. By looking at these randomly picked sentences, it is apparent that none of them is a valid description of the image. At the same time, the model prefers them over a valid, though highly underspecified, description.

There are various possible explanations for this behavior. For example, the model could be 'dazzled' by the presence of words that have a grounded referent in the image (e.g., *woman*, *girl*, or *lady* in some of the unrelated descriptions), that could lead it to assign some similarity even when the sentence is completely out of place. Conversely, the absence of words, and particularly nouns, with a clear grounded referent in the FULL description would be considered by the model as an indicator of misalignment. This could be a result of the model training data and learning objective. On the one hand, the ⟨image, text⟩ pairs scraped from the web may be poorly representative of language uses in real communicative contexts, where semantic underspecification is ubiquitous. On the other hand, the *contrastive* learning objective being employed may be too aggressive with texts that do not conform to those typically seen in training. In both cases, the similarity assigned to an underspecified description would be lower than the (possibly small) similarity assigned to an unrelated sentence with one or a few matching elements.

Moving forward Taken together, the results of the two PoCs show that CLIP struggles with semantically underspecified language. This limitation must be taken into consideration if we want to use this and similar systems to model real communicative scenarios or use them in applications that interact with human users—which is not the case for most of the tasks these models are trained and tested on. Indeed, these models may fail to retrieve an image if the language query used does not conform to the standard type of descriptions seen in training. Or, they could misunderstand inclusive uses of certain pronouns (e.g., *they*), and exhibit unwanted overspecification biases when producing an image description or referring utterance. We argue that our community, if it aims at developing language technology that can successfully and efficiently communicate with human users, should be aware of semantic underspecification and take steps toward making our models master it properly.

In the next section, we discuss how this is relevant to a range of studies exploring multimodal tasks in communicative settings.

3 Communicative Approaches to Multimodal Tasks

Mastering semantic underspecification is relevant to a wide range of studies that take a *communicative* or pragmatic approach to multimodal tasks. Below, we focus on a select sample of them[2] and discuss how they might benefit from a full mastery of the phenomenon investigated in the paper.

Image captioning with a communicative goal Standard image captioning[3] consists in generating

[2]For a recent and exhaustive survey of pragmatic work in multimodal NLP, see Fried et al. (2022).

[3]Note that, throughout this paper, we consistently avoid the term *captions* and instead use *descriptions*. We believe this terminology better reflects the fact that IC aims at generating faithful descriptions—and not captions—of images (see also

a description that is as close as possible to the content of the image. Typically, the task is not tied to a real communicative goal: image descriptions are provided by crowdworkers who are asked to mention all the important aspects of an image (Hodosh et al., 2013; Lin et al., 2014; Young et al., 2014),[4] and models are trained and evaluated to closely approximate those descriptions (Bernardi et al., 2016). To make the task more pragmatically valid, some work proposed a discriminative version of it where models need to generate a description for an image that is pragmatically informative, i.e., that is good for the image in the context of other distractor images (Andreas and Klein, 2016; Vedantam et al., 2017; Cohn-Gordon et al., 2018; Nie et al., 2020). The recent Concadia dataset (Kreiss et al., 2022b), in contrast, considers images in isolation and focuses on the communication needs of image captioning. In particular, it distinguishes between *descriptions*, useful to describe an image to someone who does not have access to the image, and *captions*, that instead complement the information of an image that is available to both interlocutors.[5]

Within both lines of research, it is to be expected that underspecified language comes into play. For example, neither the gender nor the number of people in an image may be needed to pragmatically distinguish it from other images; or, a caption complementing an image (and an accompanying text) may leave underspecified much of the information that one can get from elsewhere.[6] As such, these tasks would greatly benefit from having models opportunely dealing with this language phenomenon. In support of this—and in line with the results of our PoCs above—recent work (Kreiss et al., 2022a) showed that SotA models like CLIP are unable to account for the degree of *usefulness* of an image description, but only for its alignment. More generally, focusing on semantic underspecification of visually grounded language would be relevant to studies investigating the range of relations that texts entertain with images, including communicative goals and information needs (Kruk et al., 2019; Alikhani et al., 2020). Moreover, it would inform the *specular* task of image-to-text generation, as recently claimed by Hutchinson et al. (2022).

Goal-oriented visual question answering Standard visual question answering datasets (Antol et al., 2015) have been collected by asking crowdworkers to provide questions and answers for research purposes. In contrast, the VizWiz dataset (Gurari et al., 2018) includes questions that were asked by visually-impaired people to obtain information about visual scenes. As such, these questions are motivated by a real communicative goal and exhibit very different linguistic features compared to the questions and answers in standard datasets. For example, the questions are more ambiguous or underspecified, and the answers by the respondents are more diverse and subjective (Yang et al., 2018; Bhattacharya et al., 2019; Jolly et al., 2021). We propose that models that are equipped for semantically underspecified language should both better understand the question in relation to an image (something that current SotA models struggle with, see Chen et al., 2022) and better leverage the diversity and sparseness of the answers.

Similarly, these models may better integrate the complementary information conveyed by language and vision in, e.g., BD2BB (Pezzelle et al., 2020), a version of the visual question answering task where the correct answer (an action) results from the combination of a context (an image) and a fully *ungrounded* intention (a text); or, in other datasets that require abductive reasoning (Hessel et al., 2022). Finally, models that master semantic underspecification are expected to also deal better with related phenomena found in visual question answering, such as ambiguity and vagueness, highlighted in Bernardi and Pezzelle (2021).

Object naming and referring expressions Multimodal models should be robust to variation in object naming. For example, they should not consider as an error the use of the noun *artisan* to refer to the person in Figure 3, even if another noun, e.g., *person*, was perhaps used more frequently. At the same time, the degree of semantic specification of a naming expression should be accounted for, which would be needed to replicate patterns on human naming variation, as the ones reported by Silberer et al. (2020) and Gualdoni et al. (2022).

Naming variation is also observed in more complex visually grounded reference games, where the task is to produce a referring expression that is

a similar argument in Kreiss et al., 2022b).

[4]One exception is represented by VizWizCap (Gurari et al., 2020), a dataset of image descriptions collected with the explicit aim to be informative to visually impaired people.

[5]Captions of images in the context of news articles are a prototypical example (Hollink et al., 2016; Biten et al., 2019).

[6]It is worth mentioning that also the visual features of the scene to be described could play a non-negligible role in this sense, as recently shown by Berger et al. (2023).

pragmatically informative, i.e., that allows a listener to pick the target object (image). This task is the ideal benchmark for testing how various pragmatic frameworks, such as the Rational Speech Acts (RSA; Frank and Goodman, 2012; Goodman and Frank, 2016), can model the reference to, e.g., colors (Monroe et al., 2017) in artificial settings.

Turning to naturalistic scenarios, recent work used CLIP to quantify the properties of human referring expressions. The model was shown to capture the degree of *discriminativeness* of a referring expression over a set of images, though it assigned lower alignment scores (computed without taking into account the broader visual context) to progressively more compact utterances (Takmaz et al., 2022). Our PoCs above showed that this model conflates the semantic validity of a description with its degree of over or underspecification. However, distinguishing between the two is crucial, e.g., to assess that the expressions *the guy with the tattoos* and *the tattoo guy* are semantically equally valid, with the latter being just more underspecified (the semantic relation tying the constituents of the compound has to be inferred from the image). This can lead to models that are capable of reproducing human language patterns in certain communicative scenarios (e.g., the shortening and compression of referring utterances over an interaction, see Takmaz et al., 2020) without explicit supervision.

Visually-grounded goal-oriented dialogue All the abilities mentioned above are relevant to the development of dialogue systems that can entertain a goal-oriented conversation with human users. Examples of visually grounded goal-oriented dialogue encompass reference tasks where either *yes/no* questions (De Vries et al., 2017) or free-form, open-ended dialogue utterances (Udagawa and Aizawa, 2019; Ilinykh et al., 2019; Haber et al., 2019) are allowed to achieve a common goal, e.g., figuring out what object is being talked about or is in common between the two players. Most of these studies use datasets of interactions between human speakers to train systems that can learn to have a successful dialogue while reproducing similar linguistic and pragmatic patterns. In a few notable exceptions (Liu et al., 2018; Hawkins et al., 2020), these systems entertain an actual interaction with human users and go through a process of continual learning that leverages that *online* data. Given the communicative nature of the task, semantic underspecification is likely to be an important feature of the language used. In particular, it appears to deserve special attention when the goals involve giving and receiving visually grounded instructions (here, it is indeed one of the dimensions considered when analyzing models' results; see Kojima et al., 2021). Once again, models must be capable of dealing with semantic underspecification to communicate successfully and efficiently.

In the next section, we outline a few research directions and provide examples of concrete steps that can guide work aimed at achieving this goal.

4 Research Directions

4.1 Definitions and Operationalizations

As discussed in Section 1, semantic underspecification can be generally defined as the lack, in a linguistic signal, of part of the semantic information required to understand the message, which is typically obtained from other linguistic and non-linguistic sources. To tackle the problem at a computational level, it is important to formally define and operationalize the phenomenon. For example, by identifying which linguistic phenomena, words, or classes of words are considered by the linguistic theory as instances of semantic underspecification and under which circumstances (top-down approach). Or, by means of a data-driven measure, such as the applicability of a text to a more or less large number of visual contexts (bottom-up approach). In either case, computational methods can be used to refine or validate such definition (this is the approach used, for example, by a recent work testing the Uniform Information Density theory using language models; Giulianelli et al., 2021). Moreover, computational methods may be used to distinguish between instances of underspecification that are *deliberate* (e.g., using the pronoun *they* to refer to an individual) from those that may depend on contextual or situational aspects (e.g., not having access to some information or not mentioning something that is socially and culturally obvious).

4.2 Datasets and Annotations

Novel datasets or *ad hoc* annotations of existing resources can be collected to study underspecified language. These datasets can encompass the standard multimodal tasks (image captioning, visual question answering, etc.) and therefore be used as evaluation benchmarks to test existing models; or, new tasks can be proposed, including the prediction of an underspecification score, the para-

phrasing or explanation of an underspecifed sentence (or, *vice versa*, the de-overspecification of a sentence), and so on. Moreover, annotations may be collected at the sample and dataset level to investigate, for example, whether overspecified and underspecified image descriptions or referring utterances are equally good, informative, or inclusive[7] according to human speakers, how many and which non-linguistic cues are needed to understand them, which visual and communicative contexts elicit more underspecified language, and so on.

4.3 Model Training and Testing

Operationalizing and annotating semantic underspecification can be useful, in turn, for training and testing purposes. As for the former, sampling cases from a dataset with a varying degree of semantic underspecification can be helpful for training or fine-tuning models to make them more robust to any language. As for the latter, benchmarking a model with underspecified language can shed light on its generalization abilities and applicability to truly communicative scenarios. Moreover, a measure of a sample's semantic underspecification could be used as an additional learning signal for the training of foundational, task-agnostic multimodal models. Indeed, such a measure may indicate the extent to which language and vision convey redundant or complementary information, the relative importance of each modality, and the relation between the correctness and *self-sufficiency* of a sample. Finally, it may be interesting to leverage the degree of semantic underspecification as a dimension to which NLG models can adapt, e.g., to generate text that is more or less specified depending on the context, the interlocutor's needs or style, and the communicative goal of the linguistic interaction.

5 Conclusion

In this position paper, we argued that the NLP community must deal with semantic underspecification, that is, the possibility for a linguistic signal to convey only part of the information needed to understand a message. This is a ubiquitous phenomenon in human communication, that speakers deal with by quickly and effortlessly integrating non-linguistic information, e.g., from the surrounding visual context. We argued that research in multimodal NLP combining language and vision is ready to take on this challenge, given that SotA models that achieve unprecedented performance on a range of downstream tasks (image captioning, visual question answering, etc.) appear to struggle with it. We indicated several directions and concrete steps toward achieving this goal and discussed tasks and applications that would benefit from a full mastery of semantic underspecification.

On a technical level, our paper highlights the need to improve SotA models by making them robust to scenarios that may be different from those seen in training. In our case, CLIP suffers with sentences that resemble the language used in real communicative contexts, which poses a problem if we were to use it for modeling communicative tasks or embed it in user-facing applications. This general weakness of SotA models has been recently illustrated by Thrush et al. (2022). Using WinoGround, a dataset of carefully designed ⟨image, description⟩ pairs testing compositionality abilities, the authors reported chance-level performance for all the Transformer-based multimodal models they tested—including CLIP. A careful analysis of the samples by Diwan et al. (2022) revealed that the difficulties of the dataset go beyond dealing with compositionality, and include ambiguity aspects, reasoning abilities, and so on. In any case, these findings are informative of the flaws of the models and provide useful indications on which directions to take for improving them.

On a theoretical level, the ideas presented in our paper are consonant with a recent line of thought that advocates approaches that are aware of communicative and pragmatic aspects in language understanding and generation (Andreas, 2022; Fried et al., 2022; Giulianelli, 2022; Schlangen, 2022). We believe this is an exciting direction, and support a collaborative effort aimed at developing systems that can use language with a communicative goal.

Limitations

Semantic underspecification has been extensively studied in semantics, pragmatics, psycholinguistics, communication sciences, and cognitive sciences. In this position paper, we review this literature only superficially, although we are aware that a generalized and exhaustive understanding of the phenomenon necessarily requires knowledge of this previous work. We encourage the scholars working on this topic to embrace its complexity and depth.

The paper focuses on approaches, tasks, and

[7] These directions may also be relevant to the line of work exploring how to minimize biases and misrepresentations when describing images (e.g., Bennett et al., 2021).

models within multimodal NLP. As such, it almost completely neglects a discussion of semantic underspecification within text-only NLP. However, we are aware of the growing interest in the community at large for frameworks that propose and evaluate models in pragmatic or communicative contexts (Ruis et al., 2022; Andreas, 2022; Hu et al., 2023), and that some of the directions and steps that we propose could apply to text-only models (see, e.g., the recent, relevant work on large language models and ambiguity by Liu et al., 2023).

The two proofs of concept we report in the paper consider a rather narrow set of semantic underspecification phenomena, which may not be entirely representative. Moreover, the manual annotation that we perform, though consistent, does not adhere to any strict guidelines, and borderline cases are entrusted to the linguistic competence of the annotator. Finally, and more in general, these proofs of concepts are mostly intended to serve as a basis for the discussion and as an indication of patterns and trends. Therefore, future work should further and more thoroughly investigate this issue.

Acknowledgements

This paper owes much to the constant and passionate dialogue with the members of the Dialogue Modelling Group at the ILLC, particularly Mario Giulianelli, Ece Takmaz, and Alberto Testoni. A special thanks goes to Raquel Fernández for her valuable comments on a draft of the article.

References

Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal coherence modeling for caption generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6535, Online. Association for Computational Linguistics.

Jacob Andreas. 2022. Language models as agent models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5769–5779, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating accessibility and (mis) representation in image descriptions of race, gender, and disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19.

Uri Berger, Lea Frermann, Gabriel Stanovsky, and Omri Abend. 2023. A large-scale multilingual study of visual constraints on linguistic selection of descriptions. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2285–2299, Dubrovnik, Croatia. Association for Computational Linguistics.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Raffaella Bernardi and Sandro Pezzelle. 2021. Linguistic issues behind visual question answering. *Language and Linguistics Compass*, 15(6).

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4271–4280.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! Context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475.

Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19098–19107.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text

representation learning. In *European conference on computer vision*, pages 104–120. Springer.

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.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is Winoground hard? Investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Victor S Ferreira. 2008. Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and motivation*, 49:209–246.

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

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. 2022. Pragmatics in grounded language learning: Phenomena, tasks, and modeling approaches. *arXiv preprint arXiv:2211.08371*.

Steven Frisson. 2009. Semantic underspecification in language processing. *Language and Linguistics Compass*, 3(1):111–127.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Mario Giulianelli. 2022. Towards pragmatic production strategies for natural language generation tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7978–7984, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adele E Goldberg and Fernanda Ferreira. 2022. Good-enough language production. *Trends in Cognitive Sciences*.

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

H Paul Grice. 1969. Utterer's meaning and intentions. *The philosophical review*, 78(2):147–177.

H. Paul Grice. 1975. Logic and conversation. In D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson, Encino, California.

Eleonora Gualdoni, Thomas Brochhagen, Andreas Mädebach, and Gemma Boleda. 2022. Woman or tennis player? Visual typicality and lexical frequency affect variation in object naming. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer.

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.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Daniel W. Harris. 2020a. Intention recognition and its psychological underpinnings. *Unpublished book manuscript, CUNY Graduate Center*.

Daniel W. Harris. 2020b. What makes human communication special? *Unpublished book manuscript, CUNY Graduate Center*.

Robert Hawkins, Minae Kwon, Dorsa Sadigh, and Noah Goodman. 2020. Continual adaptation for efficient machine communication. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 408–419, Online. Association for Computational Linguistics.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The abduction of Sherlock Holmes: A dataset for visual abductive reasoning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 558–575. Springer.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Laura Hollink, Adriatik Bedjeti, Martin Van Harmelen, and Desmond Elliott. 2016. A corpus of images and text in online news. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1377–1382.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. To appear.

Ronghang Hu and Amanpreet Singh. 2021. UniT: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Ben Hutchinson, Jason Baldridge, and Vinodkumar Prabhakaran. 2022. Underspecification in scene description-to-depiction tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 1172–1184.

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*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Shailza Jolly, Sandro Pezzelle, and Moin Nabi. 2021. EaSe: A diagnostic tool for VQA based on answer diversity. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2407–2414, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Boaz Keysar. 2007. Communication and miscommunication: The role of egocentric processes. *Intercultural Pragmatics*, 4(1):71–84.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Noriyuki Kojima, Alane Suhr, and Yoav Artzi. 2021. Continual Learning for Grounded Instruction Generation by Observing Human Following Behavior. *Transactions of the Association for Computational Linguistics*, 9:1303–1319.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022a. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. 2022b. Concadia: Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4667–4684, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. 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 4622–4632.

Shalom Lappin. 2000. An intensional parametric semantics for vague quantifiers. *Linguistics and philosophy*, pages 599–620.

Stephen C Levinson. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. 2022. mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

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 *European conference on computer vision*, pages 740–755. Springer.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *arXiv e-prints*, pages arXiv–2304.

Bing Liu, Gökhan Tür, Dilek Hakkani-Tur, Pararth Shah, and Larry Heck. 2018. Dialogue learning with human teaching and feedback in end-to-end trainable task-oriented dialogue systems. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2060–2069.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. Pragmatic issue-sensitive image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. 2022. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online. Association for Computational Linguistics.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Laura Ruis, Akbir Khan, Stella Biderman, Sara Hooker, Tim Rocktäschel, and Edward Grefenstette. 2022. Large language models are not zero-shot communicators. *arXiv preprint arXiv:2210.14986*.

David Schlangen. 2022. Norm participation grounds language. In *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*, pages 62–69, Gothenburg, Sweden. Association for Computational Linguistics.

Julie C Sedivy, Michael K Tanenhaus, Craig G Chambers, and Gregory N Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition*, 71(2):109–147.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428.

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.

Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Michael Tomasello. 2005. *Constructing a language: A usage-based theory of language acquisition*. Harvard university press.

Sean Trott and Benjamin Bergen. 2022. Languages are efficient, but for whom? *Cognition*, 225:105094.

Takuma Udagawa and Akiko Aizawa. 2019. A natural language corpus of common grounding under continuous and partially-observable context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7120–7127.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Chun-Ju Yang, Kristen Grauman, and Danna Gurari. 2018. Visual question answer diversity. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations section

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Introduction (Section 1)

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Open access

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2

C ☑ Did you run computational experiments?

Section 2

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Not applicable. Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Not applicable. Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 2

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Not applicable. Left blank.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2.1

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 2.1

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.