Vision-Language Models Are Not Pragmatically Competent in Referring Expression Generation

Ziqiao Ma[*1,2,4] **Jing Ding**[*1,3] **Xuejun Zhang**[1] **Dezhi Luo**[2,4]
Jiahe Ding[1] **Sihan Xu**[1] **Yuchen Huang**[1] **Run Peng**[1] **Joyce Chai**[1]
[1]Computer Science and Engineering Division, University of Michigan
[2]Weinberg Institute for Cognitive Science, University of Michigan
[3]Department of Statistics, University of Michigan [4]GrowAI
🏠 https://vlm-reg.github.io/ 🤗 RefOI Dataset & Token-Level Human Feedback

Abstract

Referring Expression Generation (REG) is a core task for evaluating the pragmatic competence of vision-language systems, requiring not only accurate semantic grounding but also adherence to principles of cooperative communication (Grice, 1975). However, current evaluations of vision-language models (VLMs) often overlook the pragmatic dimension, reducing REG to a region-based captioning task and neglecting Gricean maxims. In this work, we revisit REG from a pragmatic perspective, introducing a new dataset (RefOI) of 1.5k images annotated with both written and spoken referring expressions. Through a systematic evaluation of state-of-the-art VLMs, we identify three key failures of pragmatic competence: (1) failure to uniquely identify the referent, (2) inclusion of excessive or irrelevant information, and (3) misalignment with human pragmatic preference, such as the underuse of minimal spatial cues. We also show that standard automatic evaluations fail to capture these pragmatic violations, reinforcing superficial cues rather than genuine referential success. Our findings call for a renewed focus on pragmatically informed models and evaluation frameworks that align with real human communication.

1 Introduction

Human language speakers routinely adjust their expressions based on listeners' perceptual and physical capabilities (Clark, 1996) and act cooperatively to enable efficient mutual understanding. Grice (1975) proposed four conversational principles—quantity, quality, relation, and manner—collectively known as the *Gricean maxims*. These maxims capture how people typically communicate: saying what is needed, when it is needed, and in a manner that is clear, relevant, and no more informative than required. In situated interactions, *referring expressions* are a key form of language use shaped by these pragmatic principles. Referring expressions have attracted long-standing interest since the last century (Winograd, 1972). From a linguistic perspective, interpreting and producing referring expressions is a natural language grounding problem (Fried et al., 2023; Mollo & Millière, 2023; Shi, 2024), requiring both semantic grounding, linking language to visual entities, and communicative grounding, establishing mutual agreement on the referent (Chai et al., 2018). From a practical perspective, this capability is essential for building robots (Qi et al., 2020) or generative AI models (Brooks et al., 2023; Yu et al., 2025) that can follow human instructions and engage in dialogue (Kollar et al., 2013; Thomason et al., 2015) with humans.

Computational models for understanding and generating these referring expressions have been extensively benchmarked in the vision-language community, ranging from the early corpora (van Deemter et al., 2006; Viethen & Dale, 2008; Mitchell et al., 2010) to the widely adopted RefCOCO series (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016; Liu et al., 2023a) and its recent variants (Tanaka et al., 2019; Lai et al., 2024; Chen et al., 2024a; Tang

[*] Authors contribute equally to this work.

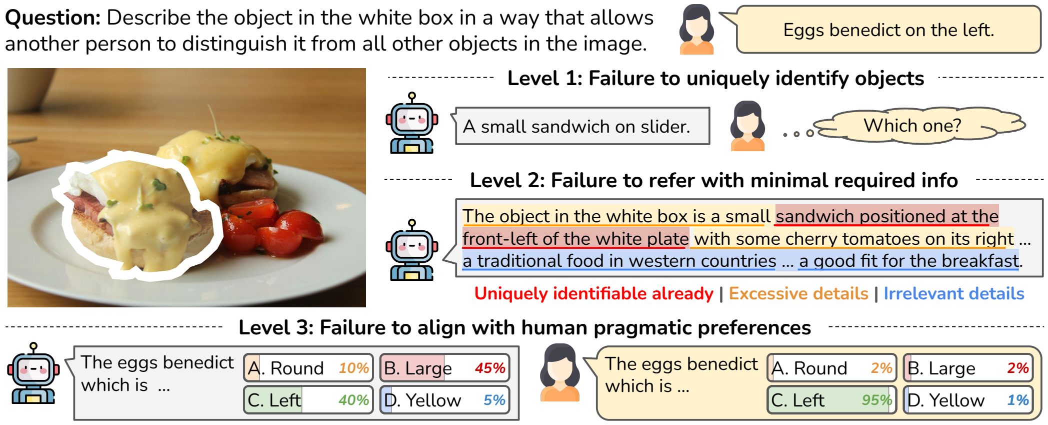

Question: Describe the object in the white box in a way that allows another person to distinguish it from all other objects in the image.
Eggs benedict on the left.
Level 1: Failure to uniquely identify objects
A small sandwich on slider.
Which one?
Level 2: Failure to refer with minimal required info
The object in the white box is a small sandwich positioned at the front-left of the white plate with some cherry tomatoes on its right ... a traditional food in western countries ... a good fit for the breakfast.
Uniquely identifiable already | Excessive details | Irrelevant details
Level 3: Failure to align with human pragmatic preferences
The eggs benedict which is ...
A. Round 10%
B. Large 45%
C. Left 40%
D. Yellow 5%
The eggs benedict which is ...
A. Round 2%
B. Large 2%
C. Left 95%
D. Yellow 1%

Figure 1: Overview of the three levels of pragmatic limitations identified in referring expression generation. While all expressions are valid regional captions under standard VLM evaluations, human-produced referring expressions are typically more concise yet still uniquely identifying. In contrast, model-generated expressions often fail to refer uniquely, include excessive or irrelevant details, and diverge from human pragmatic choices.

et al., 2024). Two core task formulations are *Referring Expression Generation (REG)*, where models generate natural language descriptions that **uniquely** identify objects in a scene, and *Referring Expression Comprehension (REC)*, where models localize the referred objects using bounding boxes or segmentation masks.

Recently, the field has seen growing excitement around vision-language models (VLMs; Team et al., 2023; OpenAI, 2024, *inter alia*) and their performance across a range of downstream tasks (Yue et al., 2024). Through instruction fine-tuning on entity-phrase mappings from text-image pairs, mechanistically grounded VLMs have been developed for fine-grained vision-language understanding at both the region (Li et al., 2022; Ma et al., 2023; Chen et al., 2023; You et al., 2023; Wang et al., 2024) and pixel level (Xia et al., 2024; Rasheed et al., 2024; Zhang et al., 2024), demonstrating strong performance on the REC task (Chen et al., 2024a). By contrast, the REG task has received significantly less attention and is often reduced to a region-based captioning task, where the requirements for unique identifiability and pragmatic conciseness, central to Gricean maxims, are largely ignored. For example, in Figure 1, while all generated expressions adequately describe the prompted region, human-produced referring expressions are much more concise yet still uniquely identifying. In contrast, model-generated expressions often fail to refer uniquely, include irrelevant information, and misalign with human language choices. This work aims to restore the pragmatic focus of the REG task as originally formulated (Dale & Reiter, 1995; 2000; Fang et al., 2013), by **evaluating whether VLMs exhibit genuine pragmatic competence**.

We introduce a new dataset consisting of around 1.5k objects within and outside the MSCOCO categories. Each object is annotated with 3 written and 2 spoken referring expressions collected from human participants. This new dataset is motivated by two main limitations of existing ones: (1) known data leakage issues in RefCOCO (Chen et al., 2020; Kamath et al., 2021), and (2) existing datasets contain LLM-generated or human-written expressions, while human communication occurs primarily through spoken language. Writing allows for planning, iterative revision, and deliberate organization, while speaking is real-time, spontaneous, and often involves formulation on the fly, which more closely reflects intuitive language use (Halliday, 1989; Peng & Harwath, 2022).

Using this dataset, we systematically evaluate a range of VLMs and identify **three levels of pragmatic limitations**: (1) **Failure to uniquely identify objects.** Many generated referring expressions are ambiguous and do not adequately disambiguate the referent, violating Grice's maxim of quality and manner. (2) **Failure to refer with minimal required information.** Models often produce excessive or irrelevant details, violating the maxims of quantity and relation. (3) **Failure to align with human pragmatic preferences.** Further analysis reveals that VLMs diverge from human pragmatic preferences, violating the maxims of manner. For example, favoring combinations of visual features over simple spatial language (Viethen & Dale, 2008; Tumu & Kordjamshidi, 2025). We also identify limitations in current evaluation metrics. Traditional heuristic metrics like BLEU (Papineni et al., 2002) fail to capture the nuanced properties of referring expressions, while using REC models as listeners (Bracha et al., 2023) risks reinforcing shortcuts that prioritize salient objects over true referential understanding. With growing efforts in building VLMs, we highlight the need for pragmatically informed models and evaluation methods that better align with human referential behavior.

2 Related Work

2.1 Pragmatics in Language Models

Pragmatics examines the contribution of context to meanings and language use. As a foundational theoretical framework, Grice (1975)'s Maxims outline a set of conversational principles that regulate how speakers contribute to effective communication, including: (1) The *Maxim of Quantity* encourages speakers to provide just the right amount of information, neither too much nor too little (Carston, 1995). (2) The *Maxim of Quality* requires that contributions be truthful and based on adequate evidence (Benton, 2016). (3) The *Maxim of Relation* demands that utterances be relevant to the ongoing discourse (Sperber & Wilson, 1986). (4) The *Maxim of Manner* urges speakers to avoid ambiguity and obscurity, and to strive for clarity and order (Koike, 1989). These maxims underpin various aspects of pragmatic reasoning in LLMs, including deixis, presupposition, indirectness, and communicative intention, assessed through tasks like interpreting indirect responses, context-dependent expressions, and dialogue turn-taking (Min et al., 2020; Hu et al., 2023; Qi et al., 2023; Sravanthi et al., 2024; Nizamani et al., 2024). When grounded to vision, these aspects are evaluated through instruction following, generation, and referring games (Zhu et al., 2021; Bao et al., 2022; Zhao et al., 2023; Nam & Ahn, 2024). We refer to Ma et al. (2025) for a comprehensive review.

2.2 Visual Grounding in Vision-Language Models

Large language models (LLMs) have demonstrated strong adaptability beyond text. In particular, a range of vision-language models (VLMs) have been developed through visual instruction tuning on paired text-image data (Liu et al., 2023b). With supervised fine-tuning using entity-phrase mappings, mechanistically grounded VLMs have achieved fine-grained vision-language understanding at both the region (Li et al., 2022; Ma et al., 2023; Chen et al., 2023; You et al., 2023; Wang et al., 2024) and pixel levels (Xia et al., 2024; Rasheed et al., 2024; Zhang et al., 2024), showing strong performance on the Referring Expression Comprehension (REC) task (Chen et al., 2024a). These models also exhibit an understanding of user-provided visual cues, giving rise to visual prompting (Yang et al., 2023a;b). Recent studies show that VLMs can interpret visual cues, such as red circles or highlights, in a zero-shot setting (Shtedritski et al., 2023; Yang et al., 2023b). This enables visual prompting through direct pixel space edits, including overlays and visual text (Li et al., 2023; Yang et al., 2023a; Lei et al., 2024; Yang et al., 2024; Wan et al., 2024). Many mechanistically grounded VLMs also incorporate explicit visual pointer tokens to represent such prompts internally (Lai et al., 2024; You et al., 2023; Zhang et al., 2024). In this work, we leverage visual prompting to reduce ambiguity when instructing VLMs to generate referring expressions.

2.3 Referring Expressions

A referring expression (RE) is a noun phrase that uniquely identifies an individual object. Referring Expression Generation (REG) is a fundamental task in pragmatic language generation that challenges models to produce such expressions. Early approaches adopt incremental algorithms (Dale & Reiter, 1995; 2000), which generate logical expressions and are closely guided by the Gricean maxims. With advances in computer vision, later work explored how visual perception shapes REs (Ren et al., 2010; FitzGerald et al., 2013; Fang et al., 2013; 2014). We refer to Krahmer & Van Deemter (2012) for a comprehensive overview of the history. More recently, a variety of large-scale referring expression datasets have been introduced. RefCLEF (Kazemzadeh et al., 2014), the RefCOCO series (Yu et al., 2016; Mao et al., 2016; Liu et al., 2023a), and Ref-L4 (Chen et al., 2024a) focus on real-world images, with varying emphases on visual attributes and linguistic complexity. Other variants extend to synthetic environments (Tanaka et al., 2019), 3D spatial contexts (Achlioptas et al., 2020; Tang et al., 2024), and reasoning-intensive domains (Lai et al., 2024). However, many of these datasets tend to emphasize the comprehension of REs (e.g., by *object detection*) over human-like language use, often featuring long and overly detailed descriptions. This trend is especially concerning in the era of VLMs, as standard REG evaluation has largely been reduced to a region-based captioning task, where models generate object descriptions conditioned on visual prompts. The core requirements of unique identifiability and pragmatic conciseness, as emphasized by the Gricean maxims, are often overlooked.

3 The RefOI Dataset

3.1 Limitations with Existing Referring Expression Datasets

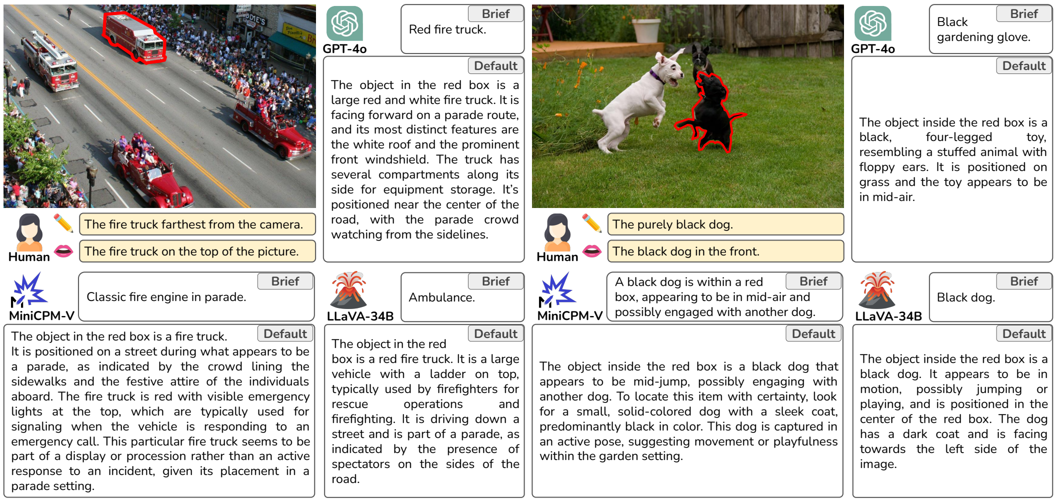
While existing datasets are available, we annotate a new one for this study, motivated by several limitations. First, prior work (Chen et al., 2020; Kamath et al., 2021) has reported data leakage in RefCOCO(+/g), as its validation and test sets overlap with the MSCOCO training set. Since MSCOCO is commonly used to train VLMs, evaluating on RefCOCO(+/g) raises concerns about data contamination. Second, existing datasets consist of LLM-generated or human-written expressions, whereas human communication primarily occurs through spoken language. LLM-generated expressions tend to include overly redundant details, often violating Gricean maxims and deviating from natural human pragmatics (Chen et al., 2024a). While written language is typically more concise, it allows for planning, revision, and deliberate organization. In contrast, spoken language is real-time, spontaneous, and involves on-the-fly formulation, more closely reflecting intuitive human language use for analytical purposes (Halliday, 1989; Peng & Harwath, 2022). To this end, we introduce a new dataset of 1,485 images featuring both COCO-class and non-COCO-class objects. Each image is annotated with 5 human-produced referring expressions: 3 written and 2 spoken.

3.2 Dataset Curation and Annotations

Image sources. We avoid using MSCOCO (Lin et al., 2014) as the image source for our dataset, as the validation and test splits of various tasks are scattered across the original COCO splits (see Fig. 4 in Chen et al. (2020)), increasing the risk of data leakage. Other commonly used datasets, such as ImageCLEF (Grubinger et al., 2006), primarily focus on stuff categories rather than things (Kazemzadeh et al., 2014), making them less suitable for evaluating object-level references. To this end, we leverage the Open Images dataset (Krasin et al., 2017; Benenson et al., 2019), which consists of deduplicated Flickr images under a CC-BY license. We use the validation split of the segmentation task, which provides high-quality pixel-level annotations for 350 entity classes, eliminating the need for manual segmentation annotation.

Balancing referent object classes We begin by mapping the 80 COCO object classes to their corresponding classes in Open Images. For each non-COCO class, we compute its cosine similarity with every COCO class and exclude those with high similarity. Among the remaining non-COCO classes, we apply incremental KMeans clustering to their Sentence-BERT embeddings (Reimers & Gurevych, 2019) to incrementally select classes that are both semantically distant from COCO classes and diverse among chosen ones. This ensures broad semantic coverage beyond COCO concepts. A key limitation of existing datasets is severe class imbalance, for example, an overwhelming focus on the *person* class. To address this, we select 25 classes from both COCO and non-COCO categories, each with 30 images. For each class, 10 masks are taken from images containing a single object instance, while the remaining 20 masks are drawn from images with multiple instances of that class. Among eligible non-COCO classes, we prioritize those that are semantically distant from COCO classes based on the clustering results. We also manually filter out low-quality and potentially harmful images. The final dataset contains 1,485 images, with 744 COCO-class masks and 741 non-COCO-class masks.

Annotating written and spoken referring expressions. Using an interactive interface (see Figure 9 in the Appendix for details), we display each mask as a bounding box on the



Human (written): The fire truck farthest from the camera.

Human (spoken): The fire truck on the top of the picture.

GPT-4o, Brief: Red fire truck.

GPT-4o, Default: The object in the red box is a large red and white fire truck. It is facing forward on a parade route, and its most distinct features are the white roof and the prominent front windshield. The truck has several compartments along its side for equipment storage. It's positioned near the center of the road, with the parade crowd watching from the sidelines.

MiniCPM-V, Brief: Classic fire engine in parade.

MiniCPM-V, Default: The object in the red box is a fire truck. It is positioned on a street during what appears to be a parade, as indicated by the crowd lining the sidewalks and the festive attire of the individuals aboard. The fire truck is red with visible emergency lights at the top, which are typically used for signaling when the vehicle is responding to an emergency call. This particular fire truck seems to be part of a display or procession rather than an active response to an incident, given its placement in a parade setting.

LLaVA-34B, Brief: Ambulance.

LLaVA-34B, Default: The object in the red box is a red fire truck. It is a large vehicle with a ladder on top, typically used by firefighters for rescue operations and firefighting. It is driving down a street and is part of a parade, as indicated by the presence of spectators on the sides of the road.

GPT-4o, Brief: Black gardening glove.

GPT-4o, Default: The object inside the red box is a black, four-legged toy, resembling a stuffed animal with floppy ears. It is positioned on grass and the toy appears to be in mid-air.

Human (written): The purely black dog.

Human (spoken): The black dog in the front.

MiniCPM-V, Brief: A black dog is within a red box, appearing to be in mid-air and possibly engaged with another dog.

MiniCPM-V, Default: The object inside the red box is a black dog that appears to be mid-jump, possibly engaging with another dog. To locate this item with certainty, look for a small, solid-colored dog with a sleek coat, predominantly black in color. This dog is captured in an active pose, suggesting movement or playfulness within the garden setting.

LLaVA-34B, Brief: Black dog.

LLaVA-34B, Default: The object inside the red box is a black dog. It appears to be in motion, possibly jumping or playing, and is positioned in the center of the red box. The dog has a dark coat and is facing towards the left side of the image.

Figure 2: Qualitative comparison of human and model referring expressions under Default and Brief prompts. Human expressions—especially in spoken form—tend to be concise and spatially grounded. In contrast, model outputs under Default prompts are often overly verbose, while Brief prompts reduce length but may omit pragmatically significant cues.

image and ask users to either type or record a referring expression, with spoken recordings automatically transcribed to text. Annotators are instructed to provide descriptions precise enough for an independent observer to identify the object unambiguously when given both the description and the original image. This process yields three written annotations and two spoken annotations for each object mask.

4 Experiments

4.1 Experiment Setups

VLMs baselines. To cover a variety of VLMs with different capabilities and training approaches, we evaluate the following models:

- VLMs build from supervised instruction fine-tuning: LLaVA-v1.5-7B/13B (Liu et al., 2023b), InternLM-XComposer-v2-7B (Dong et al., 2024);
- VLMs with both supervised fine-tuning and reinforcement learning alignment: MiniCPM-Llama3-V-v2.5-8B (Hu et al., 2024; Yu et al., 2024);
- Mechanistically grounded VLMs: GLaMM-7B (Rasheed et al., 2024) and CogVLM-Grounding-17B (Wang et al., 2024);
- Closed sourced state-of-the-art (SOTA) VLMs: GPT-4o (OpenAI, 2024).

Instructing VLMs to generate referring expressions. Our initial experiments show that VLM outputs tend to be long and verbose. To support different analytical objectives, we design two types of task prompts to guide VLMs in generating referring expressions:

- **Default (Dft.):** We follow the instruction design of referential dialogue from Zhang et al. (2024), explicitly prompting models to generate natural language descriptions that uniquely identify objects in a scene given a visual prompt.
- **Brief (Brf.):** In addition to the default prompts, we explicitly instruct VLMs to be as concise as possible, provided the object remains uniquely identifiable.

For each prompt type, we construct 10 textual variants (see Appendix A.1). During inference, we independently evaluate both the default and concise settings. For each referring task, we randomly sample one textual variant from the corresponding prompt type. For visual prompt design, we follow the practice in Chen et al. (2024b): Mechanistically grounded VLMs accept visual prompts through their dedicated pointer tokens to encode regional information. For other VLMs, we overlay visual prompts directly on the image using a red bounding box (width: 2) to ensure contrast and visibility (Shtedritski et al., 2023).

Model	Instr.	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	METEOR	CIDEr	SPICE	BERT	CLIP	REC	Human	Irrel%
LLaVA-7B	Dft.	13.27	1.60	18.09	16.30	19.29	2.10	10.50	85.51	79.02	32.41	39.46	87.30
	Brf.	28.74	6.05	**36.46**	35.50	19.15	10.80	24.59	89.02	70.72	25.51	30.57	41.95
LLaVA-13B	Dft.	8.17	1.07	11.98	10.94	16.89	0.77	7.92	84.61	79.85	30.13	46.40	91.85
	Brf.	28.96	5.81	36.44	**35.64**	20.13	8.14	21.63	88.42	72.99	28.92	32.53	49.65
LLaVA-34B	Dft.	6.29	0.78	9.82	9.11	16.15	0.07	7.61	84.39	79.86	33.42	46.53	92.90
	Brf.	28.55	6.38	32.99	31.67	20.48	9.60	16.50	88.50	74.95	35.24	36.77	56.11
XComposer	Dft.	5.25	0.65	8.38	7.81	14.58	3.10	6.37	84.11	79.86	38.06	52.19	92.81
	Brf.	13.59	2.17	17.77	16.69	19.95	5.52	10.63	85.52	79.66	38.47	51.65	80.36
MiniCPM-V	Dft.	6.38	0.67	9.86	8.78	15.28	0.05	6.30	84.29	80.38	37.93	45.12	92.97
	Brf.	16.03	3.15	19.56	18.19	18.77	6.36	11.16	86.29	78.55	35.04	45.79	72.87
GLaMM	Dft.	15.01	3.32	16.69	16.29	11.49	9.08	3.90	86.42	58.26	5.78	3.84	74.68
	Brf.	18.46	4.45	20.92	20.46	14.18	10.48	4.44	86.65	58.60	5.72	4.85	70.52
CogVLM	Dft.	31.13	**8.70**	33.89	32.32	23.50	**41.62**	24.09	89.78	66.54	33.29	26.67	**26.39**
	Brf.	**31.39**	8.69	34.70	32.94	**24.87**	41.41	**24.74**	**90.00**	69.15	38.80	33.53	29.88
GPT-4o	Dft.	7.47	0.85	11.61	10.43	17.39	0.03	7.21	84.57	**80.81**	41.29	**59.80**	89.81
	Brf.	25.30	5.78	28.76	27.36	19.02	8.17	15.31	88.11	76.58	40.08	51.72	52.75
Human	Spk.	66.18	22.58	70.15	66.45	48.28	112.04	42.35	93.89	71.60	64.56	92.20	9.15
	Wrt.	-	-	-	-	-	-	-	-	70.43	63.69	89.29	7.29

Table 1: Main results. We compare model performance under different Instr. (Instruction) settings: Dft. (Default) prompt and Brf. (Brief) prompt. All model predictions are evaluated against Human Wrt. (Written) results as the reference texts. We also compute Human Spk. (Spoken) data in comparison with human-written data. Irrel% refers to the percentage of irrelevant words in the expressions that uniquely identify an object.

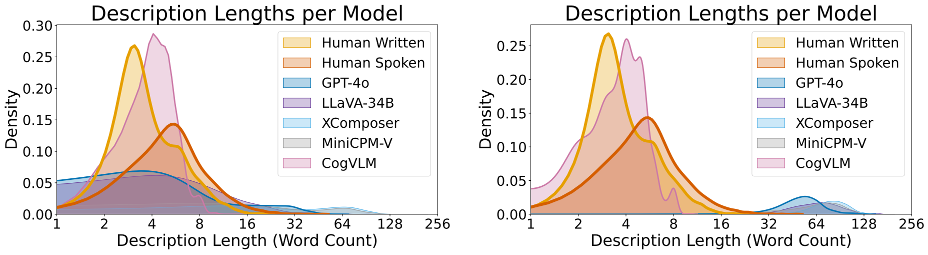
Evaluation metrics. Most of the existing automatic evaluation methods include heuristic metrics or using referring expression comprehension (REC) models as listeners (Bracha et al., 2023). Below, we summarize the main types of metrics used, including human evaluations.

- **N-gram overlap metrics:** Measure surface-level similarity between generated and reference expressions based on n-gram overlap. Includes BLEU-(1/4) (Papineni et al., 2002), ROUGE-(1/L) (Lin, 2004), and METEOR (Banerjee & Lavie, 2005).
- **Multimodal composite metrics:** Designed for image-text tasks, combining linguistic and visual context. Includes CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016).
- **Model-based metrics:** Use pretrained model embeddings to compute semantic similarity. Includes BERTScore (Zhang et al., 2020) and CLIPScore (Hessel et al., 2021).
- **Listener-based metrics:** Evaluate referential success by testing whether a listener model or human can correctly resolve the expression as an REC task (Bracha et al., 2023). We use CogVLM-Grounding (Wang et al., 2024) as the REC model as it is the reported SOTA in Chen et al. (2024a), and also collect human listers' evaluations. We compute the intersection-over-union (IoU) between the predicted bounding box and the ground truth, and consider a prediction accurate if the IoU exceeds 0.5.

Human evaluation. We provide an interactive interface (see Figure 10 in the Appendix for details) for human evaluations. We display the image without any visual cues indicating the target object, present users with a referring expression, and ask them to identify the referent. Users have three response options: (1) indicate that they cannot locate the object, (2) indicate that there are multiple potential matches, or (3) click on the image to indicate their guess. We record the pixel deviation from the nearest point on the correct object mask and consider a guess correct if the click falls within the mask. We further provide token-level human feedback by annotating text spans that aid in locating the referent when one is uniquely identifiable, and calculating the percentage of irrelevant or redundant words in each referring expression.

4.2 Evaluating Object Identifiability

In this section, we address the research question: **How well can VLMs generate referring expressions that uniquely identify an object?** We present our results in Table 1, where the last two columns represent the gold standard evaluation provided by human annotators. Human evaluators achieve high accuracy when assessing human-generated referring expressions, with 92.20% accuracy for spoken expressions and 89.29% for written ones, confirming the quality of annotations in RefOI. In contrast, all tested VLMs exhibit a substantial performance gap compared to human performance. The GPT-4o model achieves the highest accuracy among tested models with 59.80% under default prompting. Among open-source models, InternLM-XComposer-v2 performs best, reaching 52.19% accuracy.



(a) "Brief" prompts. (b) "Default" prompts.

Figure 3: Length distribution of referring expressions generated by the model using the two different prompts, including a comparison with human-written and spoken trials.

Model	Instr.	Listener Compare			Error Breakdown			Class Breakdown			Class Co-occurrence		
		Human	REC	Agree	Wrong%	Multi.%	No-Mat%	COCO	No-COCO	Δ_{Acc}	Coocc.	No-Coocc.	Δ_{Acc}
LLaVA-7B	Dft.	39.46	32.41	65.84	14.62	40.40	5.52	41.26	37.65	-3.61	18.63	81.50	-62.87
	Brf.	30.57	25.51	71.62	10.23	52.26	6.94	31.18	29.96	-1.22	10.37	71.34	-60.97
LLaVA-13B	Dft.	46.40	30.13	65.10	26.26	26.20	1.14	45.70	47.10	1.40	28.80	81.91	-53.11
	Brf.	32.53	28.92	67.99	10.30	56.63	0.54	33.47	31.58	-1.89	10.67	76.63	-65.96
LLaVA-34B	Dft.	46.53	33.42	62.14	18.72	31.52	3.23	48.25	44.80	-3.45	29.41	81.10	-51.69
	Brf.	36.77	35.24	65.03	7.34	51.45	4.44	38.04	35.49	-2.55	15.11	80.59	-65.48
XComposer	Dft.	52.19	38.06	66.11	20.20	24.92	2.69	56.05	48.31	-7.74	37.56	81.70	-44.14
	Brf.	51.65	38.47	64.09	14.28	31.45	2.62	55.78	47.50	-8.28	35.55	84.15	-48.60
MiniCPM-V	Dft.	45.12	37.93	66.38	15.75	34.55	4.58	47.98	42.24	-5.74	26.49	82.72	-56.23
	Brf.	45.79	35.04	63.62	12.19	38.99	3.03	49.46	42.11	-7.35	26.99	83.74	-56.75
GLaMM	Dft.	3.84	5.78	93.61	7.33	15.29	73.54	4.30	3.37	-0.93	1.31	8.94	-7.63
	Brf.	4.85	5.72	93.34	8.49	14.07	72.59	4.30	5.40	1.10	1.31	11.99	-10.68
CogVLM	Dft.	26.67	33.29	73.30	2.89	47.34	23.10	27.96	25.37	-2.59	13.39	53.46	-40.07
	Brf.	33.53	38.80	68.53	2.96	52.53	10.98	34.81	32.25	-2.56	16.72	67.48	-50.76
GPT-4o	Dft.	59.80	41.29	62.00	11.98	24.04	4.18	63.31	56.28	-7.03	48.14	83.33	-35.19
	Brf.	51.72	40.08	63.01	10.97	31.52	5.79	54.84	48.58	-6.26	37.36	80.69	-43.33
Human	Spk.	92.20	64.56	64.96	6.93	0.74	0.13	92.07	92.58	0.51	91.74	93.50	-1.76
	Wrt.	89.29	63.69	63.69	7.68	2.36	0.67	89.52	89.07	-0.45	88.31	91.26	-2.95

Table 2: Main results breakdown. Listener Compare: The human evaluation accuracy with REC (the evaluation result from CogVLM-Grounding) and computes Agree (the agreement between the two listeners). Error Breakdown: The percentages of three types of errors: Wrong refers to a failed guess, Multi. refers to multiple potential matches, and No-Mat refers to cases where no object can be located. Class Breakdown: The accuracy of COCO-class objects with non-COCO-class objects. The metric Δ_{Acc} shows the accuracy drop between the two categories. Class Co-occurrence: The accuracy of Coocc. images (images containing more than one object of the same class) with No-Coocc. images (images containing only one object of its class). Δ_{Acc} denotes the accuracy drop between these two categories.

Unreliable Metrics. We highlight concerns regarding the reliability of existing automatic evaluation metrics. Notably, the SOTA GPT-4o model fails to outperform other models according to N-gram overlap metrics and multimodal composite metrics. Although CogVLM-Grounding achieves the highest or near-highest scores on most automatic metrics with brief prompts, it attains only 33.53% accuracy in human evaluation. This raises concerns that current visually grounded training only enhances semantic alignment between text and image, but does not necessarily improve communicative grounding. Model-based metrics such as BERTScore and CLIPScore are ineffective at distinguishing between models; moreover, human-generated referring expressions receive relatively low CLIPScore values. While using an REC model as a listener appears to align with human evaluations for state-of-the-art models, it fails to differentiate open-source models accurately. In Table 2, we evaluate the agreement between CogVLM-Grounding's predictions and human judgments on whether a VLM-generated referring expression uniquely identifies an object, and find only 69.48% agreement on average. We further analyze why these automatic evaluation metrics fall short in pragmatic generation tasks in Section 5.2.

Breakdown. We further provide a detailed breakdown of the human evaluation results in Table 2. We analyze the reasons behind human failure to identify the referent and find that, in most cases, models fail because their descriptions lead human annotators to identify multiple possible matches. We thus divide the images based on whether they contain a single object instance or multiple instances of the same class. Most models perform reasonably well on objects without co-occurring instances of the same class, with many

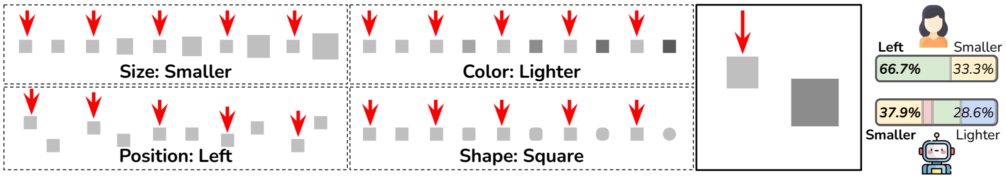
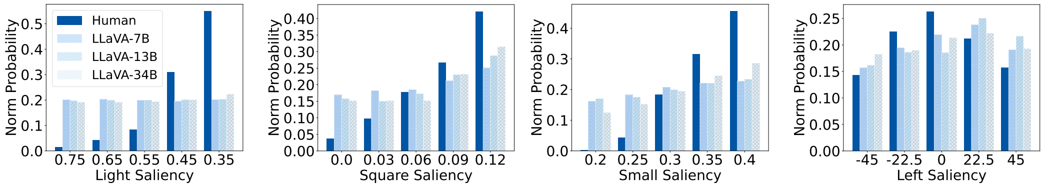


Figure 4: Synthetic dataset. Left: The gradient manipulation for each visual feature (size, color, position, shape), where the target object (red arrow) remains constant while the distractor varies along a single dimension. Right: Example of a trial in which "left," "lighter," and "smaller" can all uniquely identify the referent. Human speakers predominantly choose the spatial descriptor, whereas the VLM prefers attribute-based expressions.



(a) Lighter color. (b) Square shape. (c) Smaller size. (d) Position to the left.

Figure 5: Attribute selection as a function of feature salience. Across all four dimensions, humans readily attend to feature salience when selecting attributes for reference, whereas VLMs exhibit weaker sensitivity.

achieving over 80% accuracy. However, performance drops significantly on images with multiple co-occurring instances, ranging from 35% to 65%. This indicates that the challenge does not lie in object recognition itself, but in the models' ability to generate pragmatically effective referring expressions when multiple candidates are present. We also observe a noticeable performance drop for objects outside the COCO classes, indicating that most VLMs are biased toward MSCOCO object categories. This bias indicates that their ability to pragmatically refer to objects in more diverse, real-world settings remains limited.

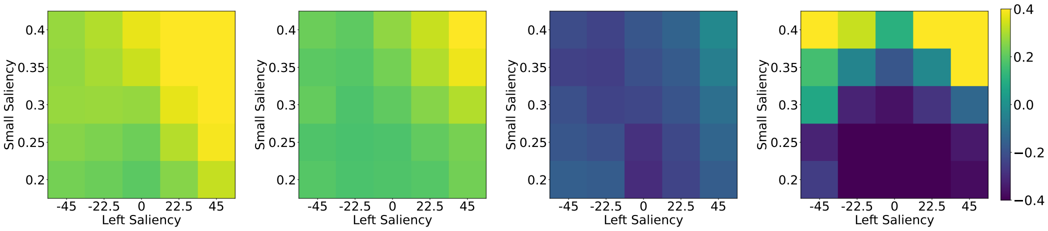
4.3 Evaluating Conciseness

In this section, we address the research question: **Can VLMs generate referring expressions with minimally required information?** In Table 1, we find that less than 10% of words in human expressions are unhelpful for disambiguating the referent. In contrast, for most VLMs, the proportion of irrelevant or redundant words can exceed 80% in default prompts and remains above 40% even when brevity is explicitly requested. CogVLM-Grounding avoids verbosity the most among identifiable referring expressions, outperforming all other models. We further present our results in Figure 3, which illustrates the distribution of referring expression lengths. Despite achieving over 90% identifiability, human-generated referring expressions are noticeably shorter than those produced by VLMs, even when models are explicitly instructed to be brief. Figure 2 provides qualitative comparisons of referring expressions. We observe that humans often rely on a single identifiable feature or a minimal combination of two features to disambiguate when necessary. In contrast, VLMs tend to produce overly detailed or irrelevant descriptions, often listing a long series of features without successfully guiding the listener to the intended object. Notably, while humans tend to rely heavily on spatial cues, VLMs are more inclined to favor combinations of visual features (Figure 14). We delve into this distinction in greater detail in Section 5.1. This issue is particularly evident in reinforcement learning-aligned models, e.g., MiniCPM-Llama3-V-v2.5, which frequently elaborate on all available details rather than prioritizing relevance and conciseness. This tendency may arise because reinforcement learning from AI feedback (Yu et al., 2024) does not accurately capture the subtleties of human pragmatics. Additional length distribution plots and qualitative examples are provided in the Appendix.

5 Further Analyses and Discussions

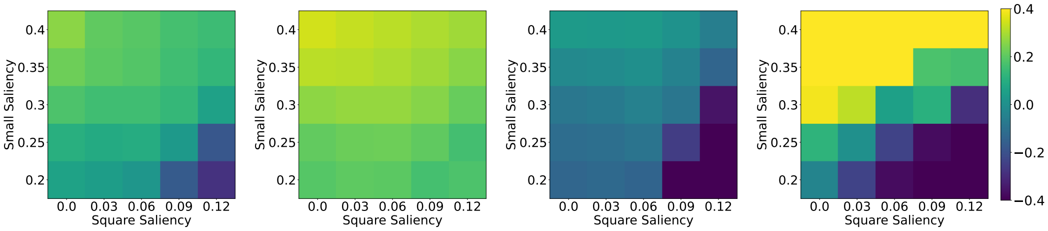
5.1 Misalignment to Human Pragmatic Preferences

In the previous section, we observe that humans tend to rely heavily on spatial cues, yet VLMs are more inclined to favor combinations of visual features. Specifically, this indicates



(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

Figure 6: Heatmap showing the difference in normalized probability of choosing "left" over "small," calculated as $\widehat{p} = \widehat{p}_{\text{small}} - \widehat{p}_{\text{left}}$. Darker colors indicate a preference for using the spatial term "left" over the size term "small."



(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

Figure 7: Heatmap showing the difference in normalized probability of choosing "square" over "small," calculated as $\widehat{p} = \widehat{p}_{\text{small}} - \widehat{p}_{\text{square}}$. Darker colors indicate a preference for using the spatial term "square" over the size term "small."

that VLMs diverge from human pragmatic preference, when multiple minimally descriptive visual features can be used to uniquely identify an object. This divergence suggests a misalignment with human pragmatic preferences, particularly in cases where multiple minimally descriptive features can uniquely identify an object. It reflects a departure from key principles of cooperative communication: selecting contextually salient attributes (Maxim of Relation) and providing just enough information for disambiguation (Quantity). We raise the question: **Do VLMs align with human pragmatic preferences?**

Synthetic data. This question is challenging to address with in-the-wild distributions due to numerous uncontrolled variables. To this end, we synthesize simple images with two co-occurring objects and task both humans and VLMs with referring to one of them. As illustrated in Figure 4, we consider four independent visual features: size, color, shape, and spatial position. The referent (marked with a red arrow) remains unchanged, while the other object is systematically altered so that the referent can always be described as one of the following expressions: *the smaller one*, *the lighter one*, *the left one*, or *the square one*. More details on data synthesis are available in Appendix B.2.

Experiment setups. To collect human pragmatic preferences, we use a similar interface where participants select the most intuitive expression for each image. Three responses are collected per image, and we average the selections to estimate the normalized probability for each expression. To evaluate VLMs' pragmatic competence rather than their performance on standard metrics, we probe the sentence probability for each expression, similar to the approach in (Zhang et al., 2025; Wang & Shi, 2025). This can be reduced to probing the token probability of the second token.

Results and discussions. In Figure 5, we plot the normalized probability of using each visual feature in referring expressions based on its visual saliency, as determined by the image synthesis setup. Further, we found that human speakers are more likely to include a particular attribute in their referring expressions as the visual salience of that feature increases. For example, when an object is especially light, small, square, or positioned to the left, the likelihood of humans mentioning those attributes rises sharply—demonstrating a gradient sensitivity to feature saliency. This aligns with the Maxim of Quantity and Manner, which jointly encourage speakers to be efficient in their referring expressions by exploiting the most disambiguating attributes available in context. In contrast, VLMs show a markedly flatter trend across salience levels, indicating that they are much less responsive to these

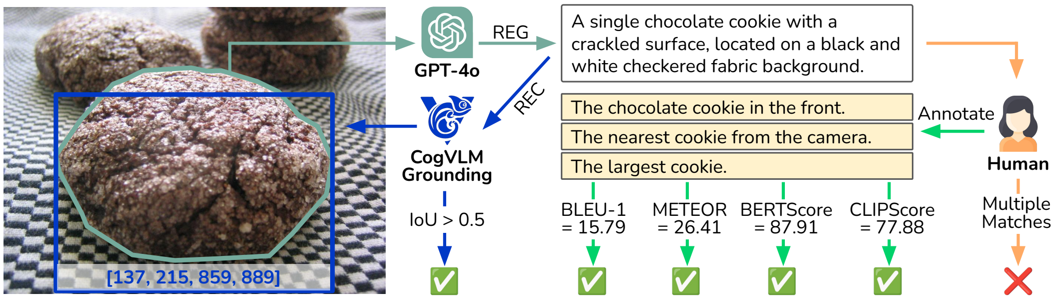

GPT-4o
REG
A single chocolate cookie with a crackled surface, located on a black and white checkered fabric background.
REC
CogVLM Grounding
IoU > 0.5
[137, 215, 859, 889]
The chocolate cookie in the front.
The nearest cookie from the camera.
The largest cookie.
Annotate
Human
Multiple Matches
BLEU-1 = 15.79
METEOR = 26.41
BERTScore = 87.91
CLIPScore = 77.88

Figure 8: A case study illustrating why automatic metrics, including heuristic measures and neural listener models, fail to accurately capture the pragmatic performance of REG.

variations. This suggests that VLMs do not internalize pragmatic principles in the same graded, context-sensitive way that humans do.

We further compare human and VLM preferences when two features coexist. In Figures 6 and 7 (with additional examples in the Appendix), we observe that human pragmatic preferences exhibit clear decision boundaries in language choice. Overall, spatial cues slightly dominate object size, as indicated by the predominance of darker cells in Figure 6d. This aligns with prior studies showing that even in simple visual scenes—where spatial language is not strictly necessary—human speakers frequently use locative expressions to refer to a target object (Viethen & Dale, 2008; Tumu & Kordjamshidi, 2025). While VLMs exhibit some human-like pragmatic preferences (e.g., comparing Figures 7(a,c,d)), they lack clear decision boundaries and often disproportionately favor one visual attribute. They also tend to omit spatial descriptors, relying instead on attribute-based constructions involving color, shape, or other non-relational features to identify the referent. This pattern supports our observation of reduced sensitivity to the Maxim of Relation and Quantity, in underspecifying information that humans include (Carston, 1995; Dale & Reiter, 1995).

5.2 Why Is (Current) Automatic Evaluation Unreliable?

We have previously identified limitations in current evaluation metrics. Here, we provide a closer analysis of why these metrics fail, supported by a case study illustrated in Figure 8.

Heuristic metrics. Heuristic metrics are not designed to account for pragmatics, and some have design flaws that are particularly undesirable for evaluating pragmatic generation. For example, BLEU is particularly ill-suited for pragmatic generation tasks. The Brevity Penalty (BP), designed to penalize overly short outputs, is undesirable when conciseness is a pragmatic virtue. For example, if the reference is "largest cookie" but the model generates "cookie," the penalty applies even if "cookie" is sufficiently identifiable. Similarly, the fragmentation penalty in METEOR targets incoherent word ordering, but in REG tasks, word order is often less important than clarity and succinctness. For instance, "the front cookie" and "cookie in the front" convey the same referent, yet the latter would be penalized despite being pragmatically acceptable.

Model-based metrics. Model-based metrics like BERTScore and CLIPScore fail to distinguish between pragmatically distinct expressions. For example, "largest cookie" and "cookie" may produce similar scores despite differing significantly in the amount of information provided. Such metrics overlook pragmatic distinctions critical to effective REG.

Neural listener models. Listener-based metrics using REC models (Bracha et al., 2023) present another problem. Only 66.81% of CogVLM-Grounding's predictions on whether a VLM-generated referring expression uniquely identifies an object align with human judgments. We observe that such a listener model often reinforces shortcuts that prioritize salient objects over genuine referential understanding. As is demonstrated in Figure 8, for instance, the speaker and listener models may succeed by associating the expression with the most visually prominent object rather than accurately resolving the intended referent.

Overall, we emphasize the urgent need to develop pragmatically-aware language generation metrics. This concern is pressing in the era of vision-language models, echoing adjacent fields like fairness and instruction generation (Zhao et al., 2023; Qiu et al., 2023; 2024).

Acknowledgments

This work was supported in part by NSF IIS-1949634, NSF SES-2128623, and the Microsoft Accelerate Foundation Models Research (AFMR) grant program. Ziqiao Ma is supported in part by the Weinberg Cognitive Science Fellowship. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies. The authors would like to thank Freda Shi, Hokin Deng, Jung-Chun Liu, Yidong Huang, Yao-An Yang for their valuable feedback. We thank all the annotators for their contributions.

Ethics Statement

Licenses. We strictly adhere to protocols governing the academic use of all research artifacts, including images, codebases, and VLMs. Our experiments use images registered under a CC-BY license rather than real user data, and we conducted a round of manual reviews to ensure there are no concerns related to privacy, bias, or societal impact.
Human study. We involved human subjects to annotate referring expressions and collect human evaluations. The institution's Institutional Review Board (IRB) deemed this project exempt from ongoing review under eResearch ID HUM00234647. While our interface includes a speech-to-text module, we did not store audio files, and no personally identifiable information was recorded.
Societal impact. For broader social impact, we aim to restore the pragmatic focus of the referring expressions task as originally formulated by evaluating whether VLMs exhibit genuine pragmatic competence. Our benchmark is designed for analytical purposes only, focusing exclusively on inference rather than training large-scale models. Consequently, the environmental impact of our work is minimal.

References

Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 422–440. Springer, 2020.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398. Springer, 2016.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2829–2842, 2022.

Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11700–11709, 2019.

Matthew A Benton. Gricean quality. *Noûs*, 50(4):689–703, 2016.

Lior Bracha, Eitan Shaar, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Disclip: Open-vocabulary referring expression generation. In *34th British Machine Vision Conference*, 2023.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

Robyn Carston. Quantity maxims and generalised implicature. *Lingua*, 96(4):213–244, 1995.

Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, volume 7, pp. 2–9, 2018.

Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. *arXiv preprint arXiv:2406.16866*, 2024a.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44393–44418, 2024b.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pp. 104–120. Springer, 2020.

Herbert H Clark. *Using language*. Cambridge university press, 1996.

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

Robert Dale and Ehud Reiter. *Building Natural Language Generation Systems*, volume 7. Cambridge University Press, 2000.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Chai. Towards situated dialogue: Revisiting referring expression generation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 392–402, 2013.

Rui Fang, Malcolm Doering, and Joyce Chai. Collaborative models for referring expression generation in situated dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

Nicholas FitzGerald, Yoav Artzi, and Luke Zettlemoyer. Learning distributions over logical forms for referring expression generation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1914–1925, 2013.

Daniel Fried, Nicholas Tomlin, Jennifer Hu, Roma Patel, and Aida Nematzadeh. Pragmatics in language grounding: Phenomena, tasks, and modeling approaches. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12619–12640, 2023.

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

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. The iapr tc-12 benchmark: A new evaluation resource for visual information systems. In *International workshop ontoImage*, volume 2, 2006.

Michael Alexander Kirkwood Halliday. *Spoken and written language*. OXford University Press, 1989.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 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 (Volume 1: Long Papers)*, pp. 4194–4213, 2023.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1780–1790, 2021.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 787–798, 2014.

Dale April Koike. Requests and the role of deixis in politeness. *Journal of Pragmatics*, 13(2): 187–202, 1989.

Thomas Kollar, Vittorio Perera, Daniele Nardi, and Manuela Veloso. Learning environmental knowledge from task-based human-robot dialog. In *2013 IEEE International Conference on Robotics and Automation*, pp. 4304–4309. IEEE, 2013.

Emiel Krahmer and Kees Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9579–9589, 2024.

Xuanyu Lei, Zonghan Yang, Xinrui Chen, Peng Li, and Yang Liu. Scaffolding coordinates to promote vision-language coordination in large multi-modal models. *arXiv preprint arXiv:2402.12058*, 2024.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975, 2022.

Zongjie Li, Chaozheng Wang, Chaowei Liu, Pingchuan Ma, Daoyuan Wu, Shuai Wang, and Cuiyun Gao. Vrptest: Evaluating visual referring prompting in large multimodal models. *arXiv preprint arXiv:2312.04087*, 2023.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 23592–23601, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems*, volume 36, 2023b.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. *arXiv preprint arXiv:2502.12378*, 2025.

Ziqiao Ma, Jiayi Pan, and Joyce Chai. World-to-words: Grounded open vocabulary acquisition through fast mapping in vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 524–544, 2023.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 11–20, 2016.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5783–5797, 2020.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. Natural reference to objects in a visual domain. In *Proceedings of the 6th international natural language generation conference*, 2010.

Dimitri Coelho Mollo and Raphaël Millière. The vector grounding problem. *arXiv preprint arXiv:2304.01481*, 2023.

Heejeong Nam and Jinwoo Ahn. Visual contexts clarify ambiguous expressions: A benchmark dataset. *arXiv preprint arXiv:2411.14137*, 2024.

Rashid Nizamani, Sebastian Schuster, and Vera Demberg. Siga: A naturalistic nli dataset of english scalar implicatures with gradable adjectives. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 14784–14795, 2024.

OpenAI. Hello gpt-4o, May 2024. URL https://openai.com/index/hello-gpt-4o/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Puyuan Peng and David Harwath. Word discovery in visually grounded, self-supervised speech models. In *Proc. Interspeech 2022*, pp. 2823–2827, 2022.

Peng Qi, Nina Du, Christopher D Manning, and Jing Huang. Pragmaticqa: A dataset for pragmatic question answering in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 6175–6191, 2023.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8358–8375, 2023.

Haoyi Qiu, Wenbo Hu, Zi-Yi Dou, and Nanyun Peng. Valor-eval: Holistic coverage and faithfulness evaluation of large vision-language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 1783–1805, 2024.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, pp. 3982–3992, 2019.

Yuan Ren, Kees Van Deemter, and Jeff Z Pan. Charting the potential of description logic for the generation of referring expressions. In *Proceedings of the 6th International Natural Language Generation Conference*, 2010.

Haoyue Freda Shi. *Learning Language Structures through Grounding*. PhD thesis, Toyota Technological Institute at Chicago, 2024.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA, 1986.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 12075–12097, 2024.

Mikihiro Tanaka, Takayuki Itamochi, Kenichi Narioka, Ikuro Sato, Yoshitaka Ushiku, and Tatsuya Harada. Generating easy-to-understand referring expressions for target identifications. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5794–5803, 2019.

Zineng Tang, Lingjun Mao, and Alane Suhr. Grounding language in multi-perspective referential communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19727–19741, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pp. 1923–1929, 2015.

Akshar Tumu and Parisa Kordjamshidi. Exploring spatial language grounding through referring expressions. *arXiv preprint arXiv:2502.04359*, 2025.

Kees van Deemter, Ielka van der Sluis, and Albert Gatt. Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the fourth international natural language generation conference*, pp. 130–132, 2006.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Jette Viethen and Robert Dale. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pp. 59–67, 2008.

David Wan, Jaemin Cho, Elias Stengel-Eskin, and Mohit Bansal. Contrastive region guidance: Improving grounding in vision-language models without training. In *European Conference on Computer Vision*, pp. 198–215. Springer, 2024.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024.

Yixuan Wang and Freda Shi. Logical forms complement probability in understanding language model (and human) performance. *arXiv preprint arXiv:2502.09589*, 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Terry Winograd. Understanding natural language. *Cognitive psychology*, 3(1):1–191, 1972.

Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023a.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. Fine-grained visual prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023b.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2023.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pp. 69–85. Springer, 2016.

Shoubin Yu, Difan Liu, Ziqiao Ma, Yicong Hong, Yang Zhou, Hao Tan, Joyce Chai, and Mohit Bansal. Veggie: Instructional editing and reasoning video concepts with grounded generation. *arXiv preprint arXiv:2503.14350*, 2025.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lingjun Zhao, Khanh Nguyen, and Hal Daumé Iii. Define, evaluate, and improve task-oriented cognitive capabilities for instruction generation models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3688–3706, 2023.

Hao Zhu, Graham Neubig, and Yonatan Bisk. Few-shot language coordination by modeling theory of mind. In *International Conference on Machine Learning*, pp. 12901–12911. PMLR, 2021.

A Reproducibility

A.1 Prompt Templates

Prompt Templates for Model REG Task

Default:

1. Describe the object in the red box in a way that allows another person to distinguish it from all other objects in the image.
2. Give a clear and specific description of the object in the red box so that another user can find it without hesitation.
3. Describe the object in the red box, so that another user can identify the object from other objects.
4. Describe the object in the red box in a way that allows another person to find that one specific object.
5. Provide a description of the object in the red box so that another person can recognize and identify that unique item.
6. Describe the object in the red box so that another person can pinpoint that exact object among others.
7. Explain the features of the object in the red box that make it stand out, so another person can find it.
8. Describe the object inside the red box so that someone else can locate that particular item with certainty.
9. Describe the object in the red box so that another person can find that one specific object.
10. Give a description of the object in the red box so that another user can identify the exact unique object.

Brief:

1. Describe the red-boxed object using the fewest words while ensuring it can be uniquely identified.
2. Give a minimal description that allows someone to find the exact object in the red box.
3. Use the least words necessary to ensure the red-boxed object is unmistakably identifiable.
4. Provide a short yet precise description so the red-boxed object can be uniquely located.
5. Describe the object in the red box concisely, ensuring it is the only possible match.
6. Identify the red-boxed object using the fewest words while making it uniquely findable.
7. Give a brief but unambiguous description that guarantees the red-boxed object can be found.
8. Provide the shortest possible description that still allows precise identification of the red-boxed object.
9. Describe the red-boxed object in minimal words while ensuring no confusion with other objects.
10. Use as few words as possible to describe the red-boxed object in a way that guarantees unique identification.

Chain-of-Thought:
We append “Think step by step.” to the end of each Brief prompt.

Prompt Templates for Model REG Task

Gricean:
We append the following definition to the end of each Brief prompt.

Follow Gricean principles. Grice outlined four key categories, or maxims, of conversation—quantity, quality, relation, and manner.

The maxim of quantity is: be informative.
Submaxims:
1. Make your contribution as informative as is required (for the current purposes of the exchange). 2. Do not make your contribution more informative than is required.
In his book, Grice uses the following analogy for this maxim: "If you are assisting me to mend a car, I expect your contribution to be neither more nor less than is required. If, for example, at a particular stage I need four screws, I expect you to hand me four, rather than two or six."

The maxim of quality is: be truthful.
Supermaxim: Try to make your contribution one that is true.
Submaxims:
1. Do not say what you believe is false. 2. Do not say that for which you lack adequate evidence.
In his book, Grice uses the following analogy for this maxim: "I expect your contributions to be genuine and not spurious. If I need sugar as an ingredient in the cake you are assisting me to make, I do not expect you to hand me salt; if I need a spoon, I do not expect a trick spoon made of rubber."

The maxim of relation is: be relevant.
The information provided should be relevant to the current exchange and omit any irrelevant information.
In his book, Grice uses the following analogy for this maxim: "I expect a partner's contribution to be appropriate to the immediate needs at each stage of the transaction. If I am mixing ingredients for a cake, I do not expect to be handed a good book, or even an oven cloth (though this might be an appropriate contribution at a later stage)."
With respect to this maxim, Grice writes, "Though the maxim itself is terse, its formulation conceals a number of problems that exercise me a good deal: questions about what different kinds and focuses of relevance there may be, how these shift in the course of a talk exchange, how to allow for the fact that subjects of conversations are legitimately changed, and so on. I find the treatment of such questions exceedingly difficult, and I hope to revert to them in later work."

The maxim of manner is: be clear.
Whereas the previous maxims are primarily concerned with *what* is said, the maxims of manner are concerned with how it is said.
Supermaxim: Be perspicuous.
Submaxims:
1. Avoid obscurity of expression — i.e., avoid language that is difficult to understand. 2. Avoid ambiguity — i.e., avoid language that can be interpreted in multiple ways. 3. Be brief — i.e., avoid unnecessary verbosity. 4. Be orderly — i.e., provide information in an order that makes sense, and makes it easy for the recipient to process it.

Prompt Template for CogVLM-Grounding REC Task

{*referring_expression*}. Please provide the bounding box in the format [[x0,y0,x1,y1]] for the object in the image.

A.2 Computational Resources.

Our experiments were conducted using 4 A40 GPUs and 1 A6000 GPU, with a total computational cost of approximately 140 A40 GPU hours.

B Dataset Details

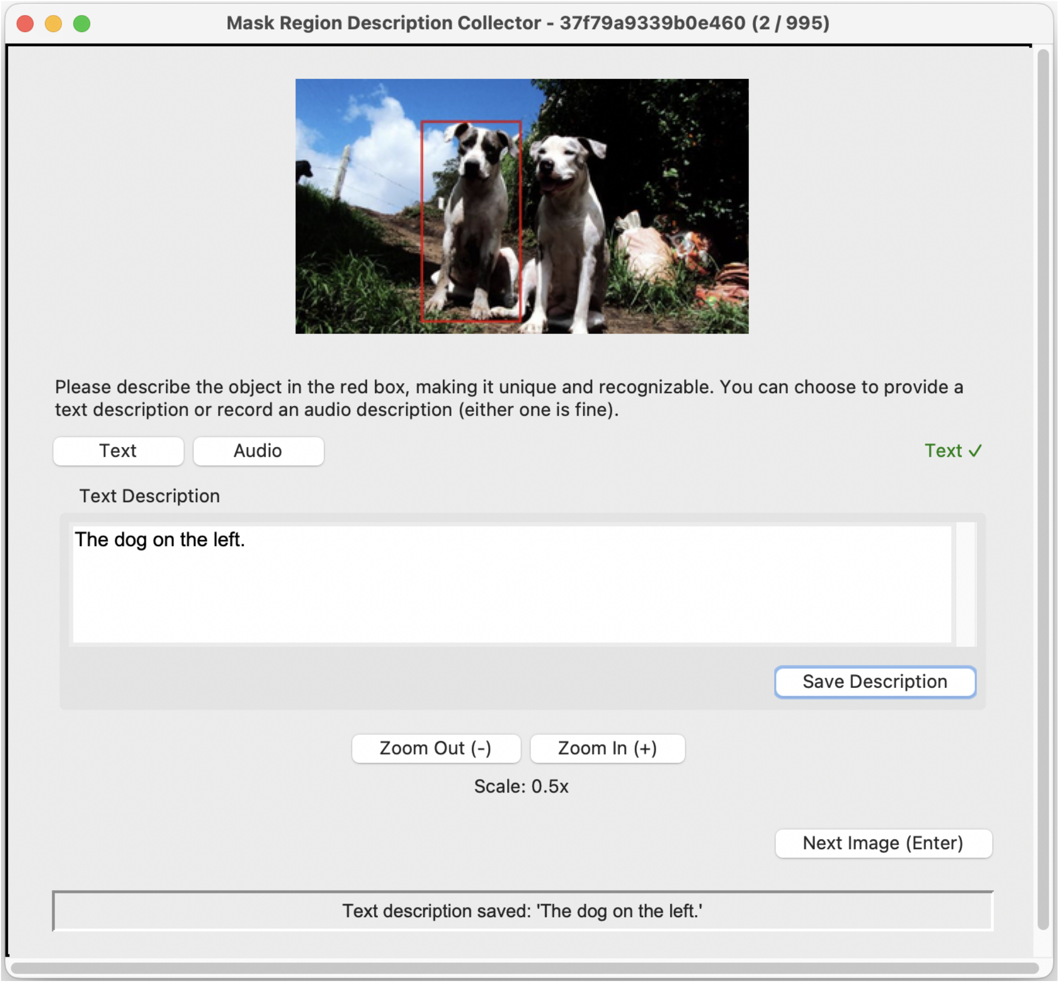
B.1 Annotator Instructions

We recruited and trained 4 annotators for generating high-quality referring expressions, 9 annotators for evaluating VLM-generated referring expressions, and 12 human subjects for collecting pragmatic preferences. To collect pragmatic preferences, we ensured that each human subject received no more than 25 examples per session. After completing a batch, participants were required to take a break of at least 10 minutes before proceeding to the next batch. We show our annotation interface in Figure 9, 10 and 11.

B.2 Synthetic Data Generation

As illustrated in Figure 4, we consider four independent visual features: size, color, shape, and spatial position. The referent (marked with a red arrow) remains unchanged, while the other object is systematically altered so that the referent can always be described as either *the smaller one*, *the lighter one*, *the left one*, or *the square one*. The referent is consistently a square with a size of 0.2 and a gray value of 0.75. The other object is a rounded square, with its attributes varying as follows, leading to 625 images:

- Size: $\{0.2, 0.25, 0.30, 0.35, 0.40\}$
- Gray value: $\{0.75, 0.65, 0.55, 0.45, 0.35\}$
- Rounding size: $\{0.00, 0.03, 0.06, 0.09, 0.12\}$
- Deviation angle from lateral direction: $\{-45^\circ, -22.5^\circ, 0^\circ, 22.5^\circ, 45^\circ\}$



Mask Region Description Collector - 37f79a9339b0e460 (2 / 995)

Please describe the object in the red box, making it unique and recognizable. You can choose to provide a text description or record an audio description (either one is fine).

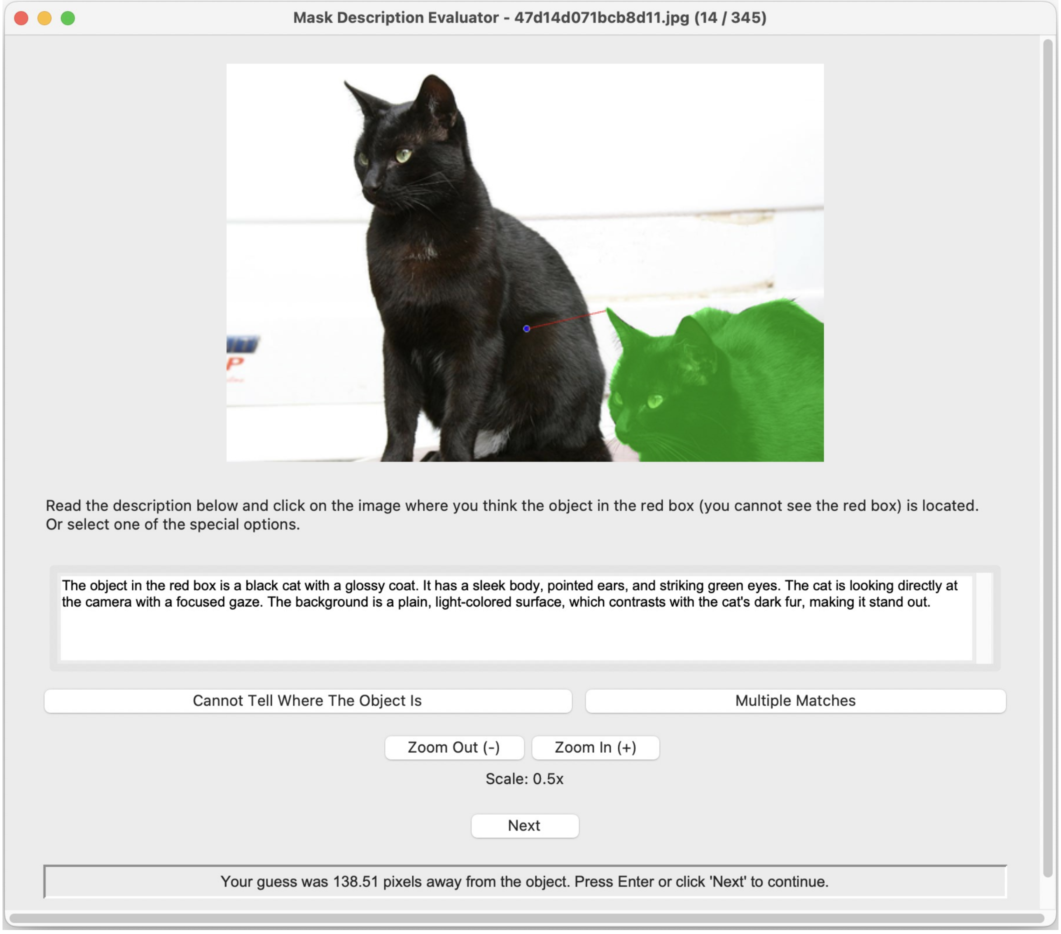
Text Description

The dog on the left.

Scale: 0.5x

Text description saved: 'The dog on the left.'

Figure 9: The annotation interface. The user is required to enter text or provide a speech-to-text description of the object within the red box, ensuring that another observer can uniquely identify the object in the image.



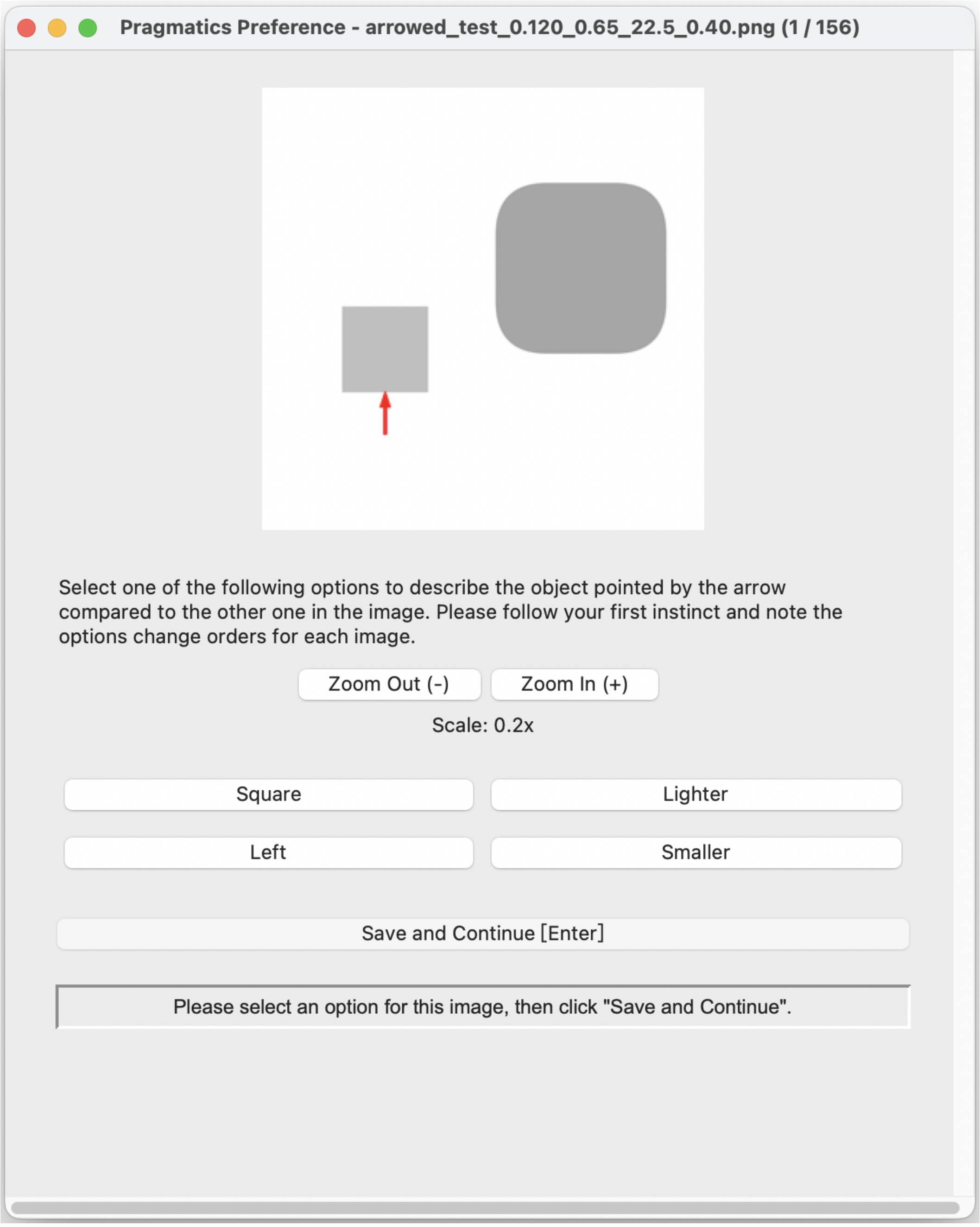
Mask Description Evaluator - 47d14d071bcb8d11.jpg (14 / 345)

Read the description below and click on the image where you think the object in the red box (you cannot see the red box) is located. Or select one of the special options.

The object in the red box is a black cat with a glossy coat. It has a sleek body, pointed ears, and striking green eyes. The cat is looking directly at the camera with a focused gaze. The background is a plain, light-colored surface, which contrasts with the cat's dark fur, making it stand out.

Your guess was 138.51 pixels away from the object. Press Enter or click 'Next' to continue.

Figure 10: The human evaluation interface. The user is required to click on the corresponding object in the image based on the given description. The interface then displays the shortest distance between the clicked point and the nearest point on the mask of the target object, which will be 0 if the click is inside the mask. If the user cannot find any object matching the description or identifies multiple possible objects, they should click the corresponding button instead, and the distance will not be computed.



Pragmatics Preference - arrowed_test_0.120_0.65_22.5_0.40.png (1 / 156)

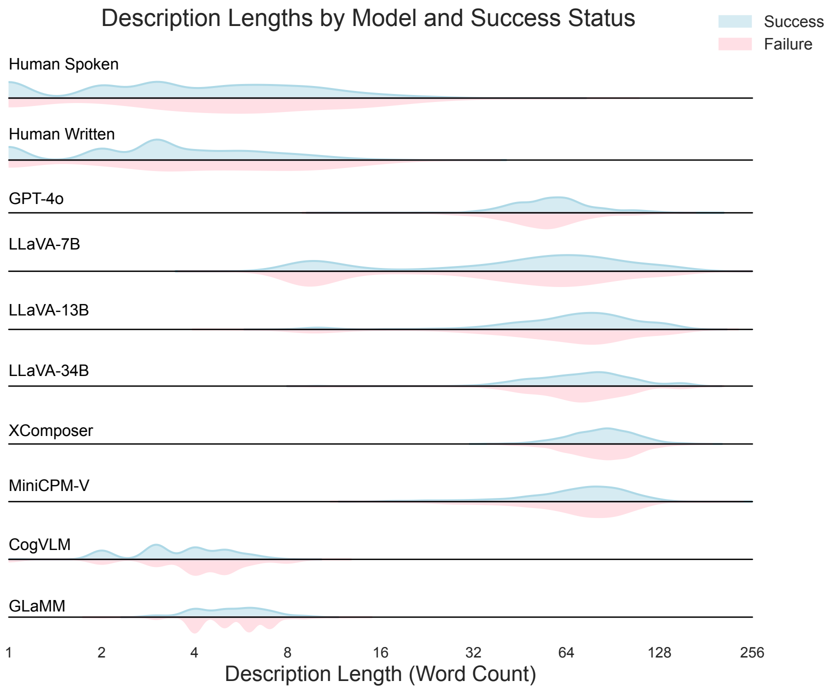
Select one of the following options to describe the object pointed by the arrow compared to the other one in the image. Please follow your first instinct and note the options change orders for each image.

Scale: 0.2x

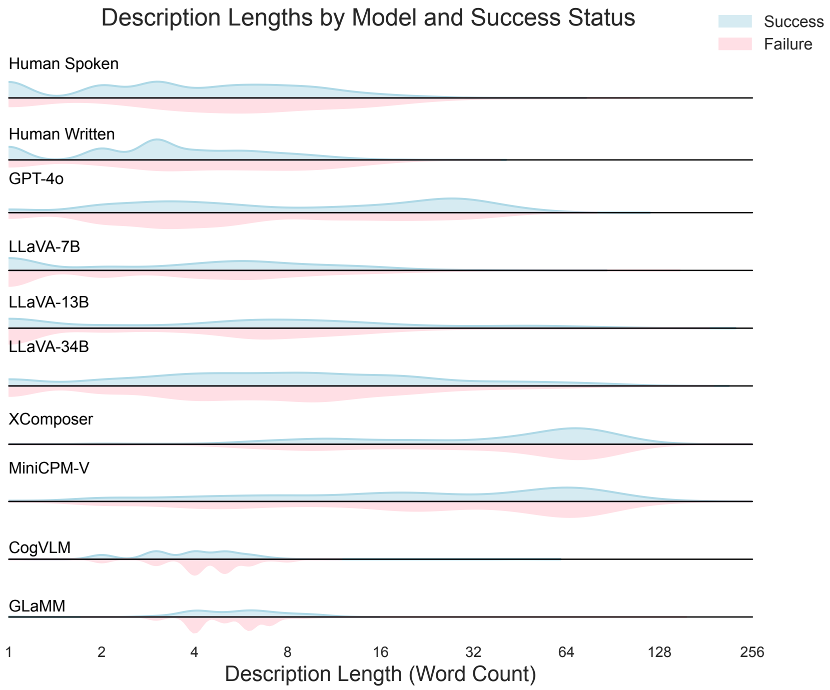
Please select an option for this image, then click "Save and Continue".

Figure 11: Interface for pragmatic preference collection. Participants are instructed to choose one of four descriptions for the object indicated by the red arrow, guided by their immediate intuition. Each session includes up to 25 examples, followed by a mandatory break of at least 10 minutes before continuing.

B.3 Length Distribution



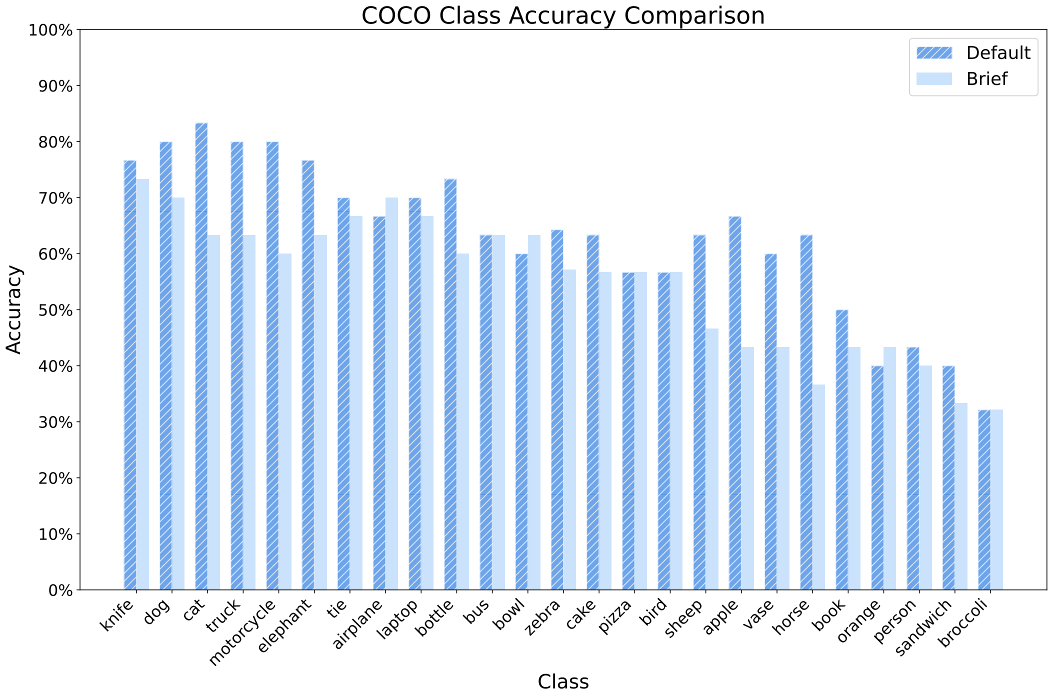
(a) "Default" prompts.



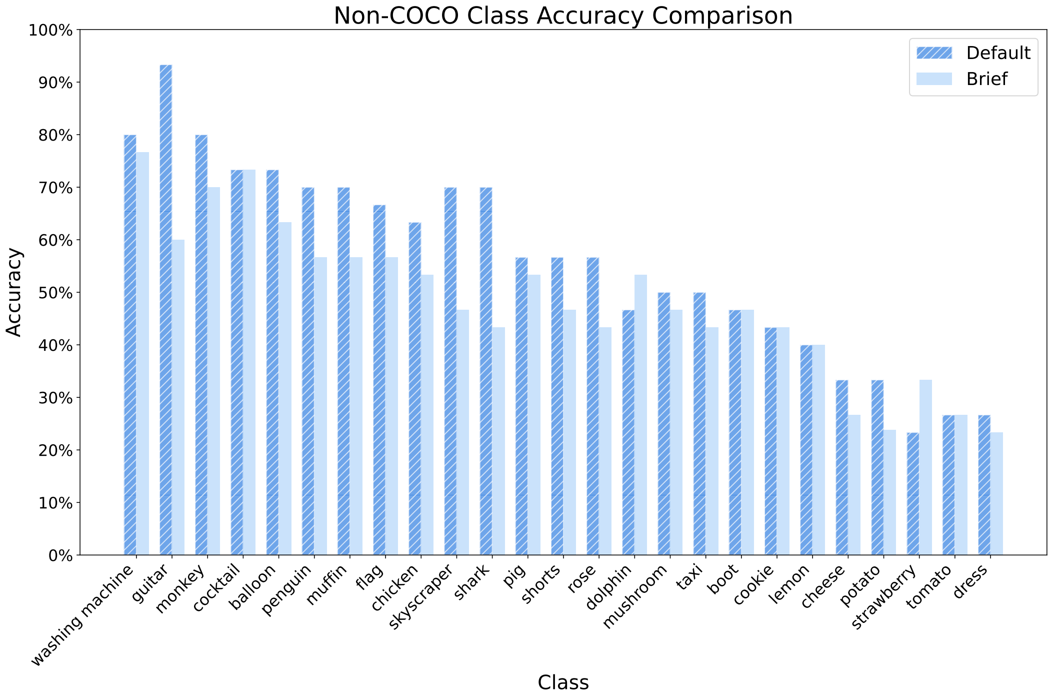
(b) "Brief" prompts.

Figure 12: Length distribution of referring expressions generated by the model using the two different prompts, including a comparison with human-written and spoken trials. We filter out empty expressions, which are not included in the length statistic for failure cases.

B.4 Accuracy by Class



(a) COCO classes.

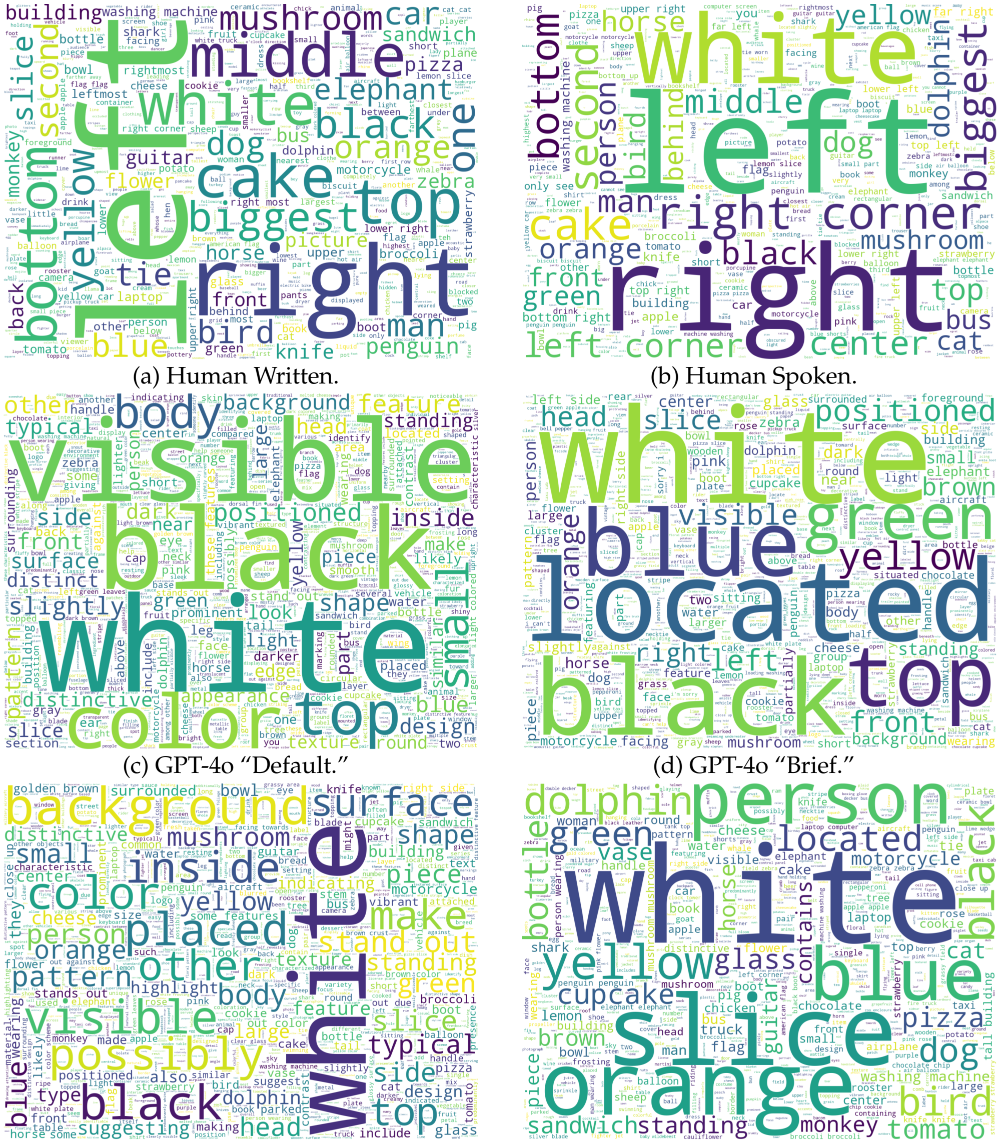


(b) Non-COCO classes.

Figure 13: Accuracy of GPT-4o under “Default” and “Brief” prompts, grouped by COCO (a) and non-COCO (b) classes and sorted by average accuracy. Overall, COCO classes achieve higher accuracy.

B.5 Word Cloud

We present all the word clouds generated by humans and different models here. While human rely heavily on spatial cues, all models favor visual features.



(a) Human Written.

(b) Human Spoken.

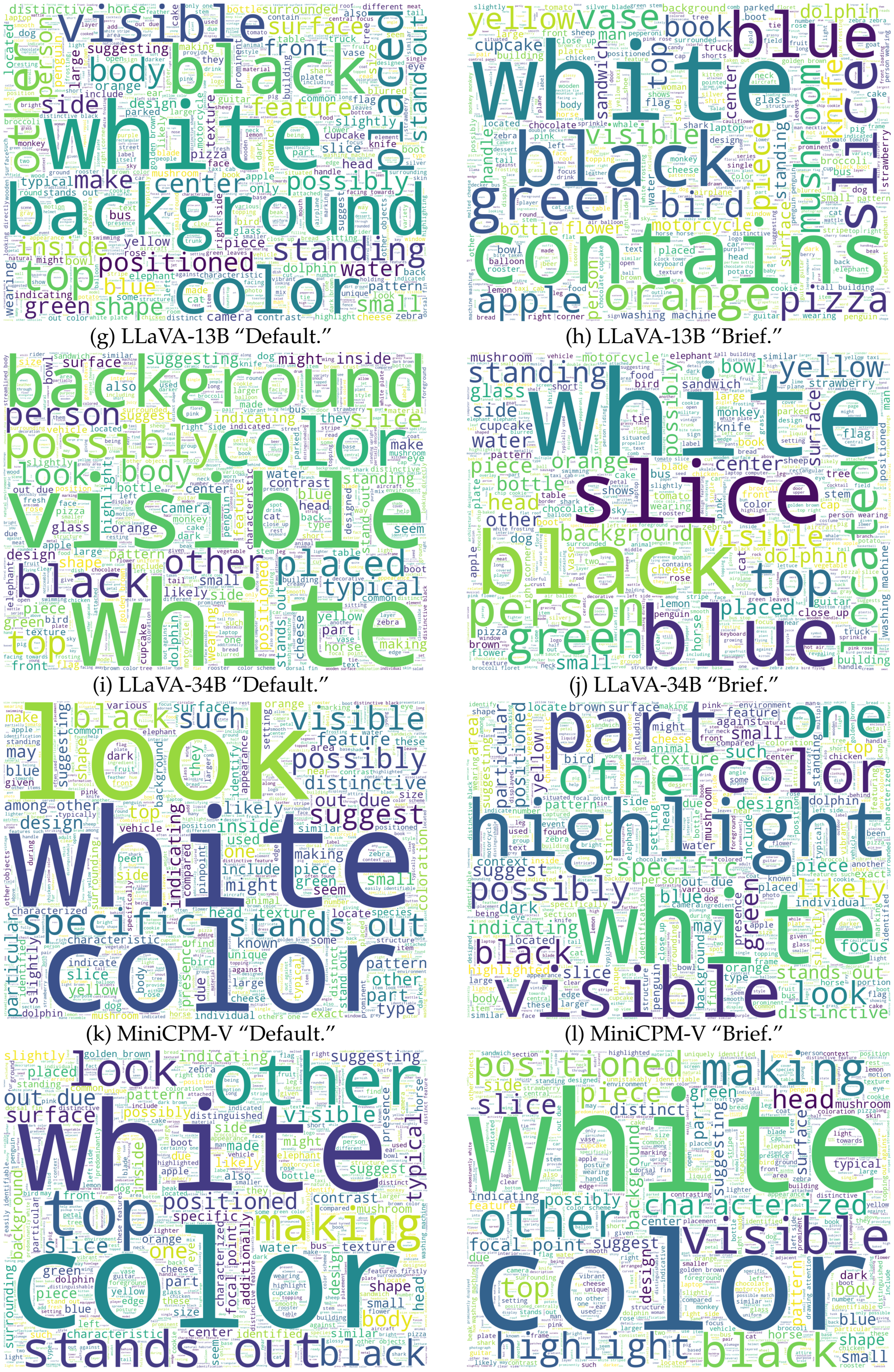
(c) GPT-4o “Default.”

(d) GPT-4o “Brief.”

(e) LLaVA-7B “Default.”

(f) LLaVA-7B “Brief.”

Figure 14: Word clouds generated by humans and different models under “Default” and “Brief” prompting conditions (1/3).



(g) LLaVA-13B "Default."

(h) LLaVA-13B "Brief."

(i) LLaVA-34B "Default."

(j) LLaVA-34B "Brief."

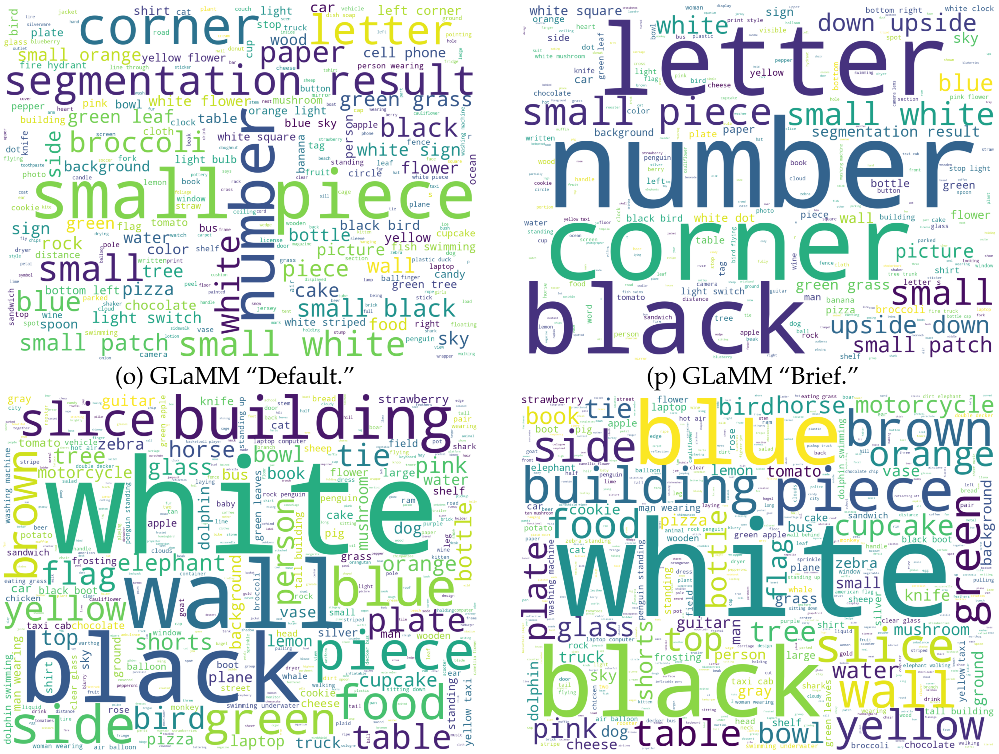
(k) MiniCPM-V "Default."

(l) MiniCPM-V "Brief."

(m) XComposer "Default."

(n) XComposer "Brief."

Figure 14: (continued) Word clouds generated by models (2/3).



(o) GLaMM “Default.”

(p) GLaMM “Brief.”

(q) CogVLM “Default.”

(r) CogVLM “Brief.”

Figure 14: (continued) Word clouds generated by models (3/3).

C Does Simple Prompting Fix Pragmatic Limitations?

C.1 Reasoning-oriented Prompts

Prior work (Wei et al., 2022) suggests that chain-of-thought reasoning can improve performance. Given the pragmatic failures of VLMs identified in our experiments, we further ask: **Can these shortcomings be mitigated through simple prompting techniques?** To investigate this, we conducted supplementary experiments on GPT-4o by evaluating two enhanced prompts that elicit better reasoning, building on our brief prompt:

- **Chain-of-Thought (CoT):** Chain-of-thought prompting (Wei et al., 2022) directs a model to *think step by step*, explicitly articulating intermediate reasoning before committing to an answer. We follow the common setting by appending the sentence "Think step by step." to our brief prompt.
- **Gricean (Grn.):** As the *Gricean maxims* provide a normative framework for cooperative communication, we explicitly instruct the model to follow the Gricean maxims and embed a definition sourced from Wikipedia.

C.2 Results and discussions

Model	Instr.	BLEU-1	BLEU-4	ROUGE-1	ROUGE-L	METEOR	CIDEr	SPICE	BERT	CLIP	REC	Human	Irrel%
GPT-4o	Dft.	7.47	0.85	11.61	10.43	17.39	0.03	7.21	84.57	**80.81**	41.29	59.80	89.81
	Brf.	**25.30**	**5.78**	**28.76**	**27.36**	19.02	8.17	**15.31**	**88.11**	76.58	40.08	51.72	**52.75**
	CoT	19.98	4.02	23.42	21.84	18.15	10.15	13.36	86.83	77.59	**46.67**	**65.59**	65.71
	Grn.	21.43	4.33	25.62	24.01	**20.35**	**10.85**	13.87	87.57	78.50	42.10	63.50	62.40
Human	Spk.	66.18	22.58	70.15	66.45	48.28	112.04	42.35	93.89	71.60	64.56	92.20	9.15
	Wrt.	-	-	-	-	-	-	-	-	70.43	63.69	89.29	7.29

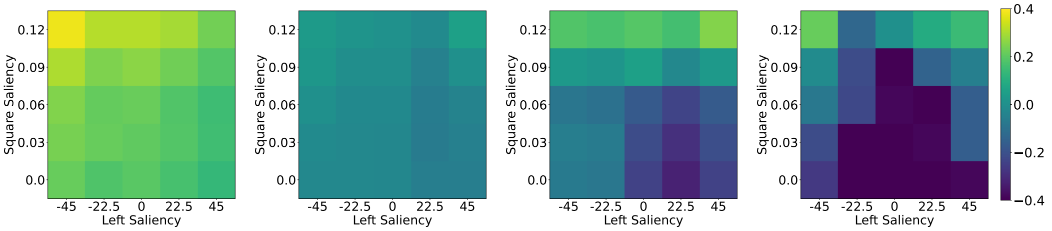
Table 3: Supplementary results for reasoning-oriented prompts of GPT-4o.

Model	Instr.	Listener Compare			Error Breakdown			Class Breakdown			Class Co-occurrence		
		Human	REC	Agree	Wrong%	Multi.%	No-Mat%	COCO	No-COCO	Δ_{Acc}	Coocc.	No-Coocc.	Δ_{Acc}
GPT-4o	Dft.	59.80	41.29	62.00	11.98	24.04	4.18	63.31	56.28	-7.03	48.14	83.33	-35.19
	Brf.	51.72	40.08	63.01	10.97	31.52	5.79	54.84	48.58	-6.26	37.36	80.69	-43.33
	CoT	65.59	46.67	63.55	14.28	17.37	2.76	68.15	63.02	-5.13	54.98	86.99	-32.01
	Grn.	63.50	42.10	60.52	14.08	19.80	2.62	65.05	61.94	-3.11	51.96	86.79	-34.83
Human	Spk.	92.20	64.56	64.96	6.93	0.74	0.13	92.07	92.58	0.51	91.74	93.50	-1.76
	Wrt.	89.29	63.69	63.69	7.68	2.36	0.67	89.52	89.07	-0.45	88.31	91.26	-2.95

Table 4: Supplementary results breakdown for reasoning-oriented prompts of GPT-4o.

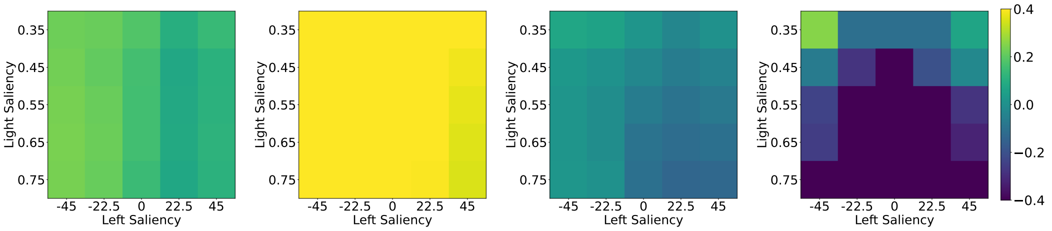
As shown in Table 3, both prompts produce improvements in human evaluation, with the CoT prompt achieving the highest accuracy. In Table 4, we also observe a reduction in the rate of multiple matches, suggesting that these prompts encourage more discriminative expressions of reference. Although our findings indicate that enhanced prompts can improve object identifiability, most of the other automatic metrics fail to reflect this improvement, exhibiting only minor fluctuations or even declines, which echoes our earlier critique that such metrics overlook pragmatic quality. For conciseness, however, we observe that the proportion of irrelevant or redundant words increases under both enhanced prompts, suggesting that the gains in object identifiability come at the cost of brevity. This highlights a trade-off: Reasoning-oriented prompts may improve referential success, but often lead to more verbose descriptions. In other words, prompt engineering alone is insufficient to achieve both accurate and concise referring expressions. Overall, the results suggest that while reasoning-oriented prompting helps to some extent, it does not address the underlying limitations that prevent models from matching human performance, which underscores the need to instill pragmatic competence at training time as an intrinsic capability, rather than rely solely on prompt engineering.

D Additional Qualitative Comparisons



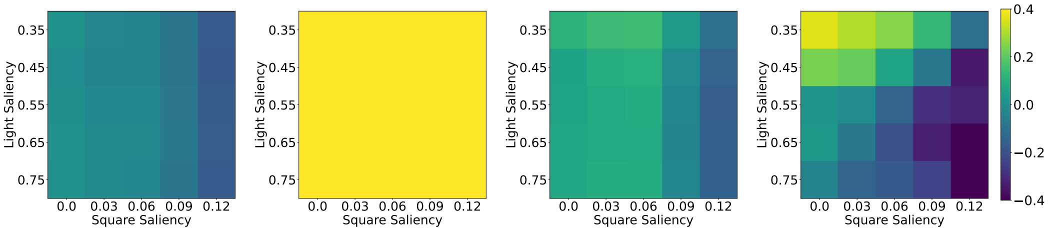
(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

Figure 15: Heatmap showing the difference in normalized probability of choosing "square" over "left," calculated as $\widehat{p} = \widehat{p}_{\text{square}} - \widehat{p}_{\text{left}}$. Darker colors indicate a preference for using the spatial term "left" over the shape term "square."



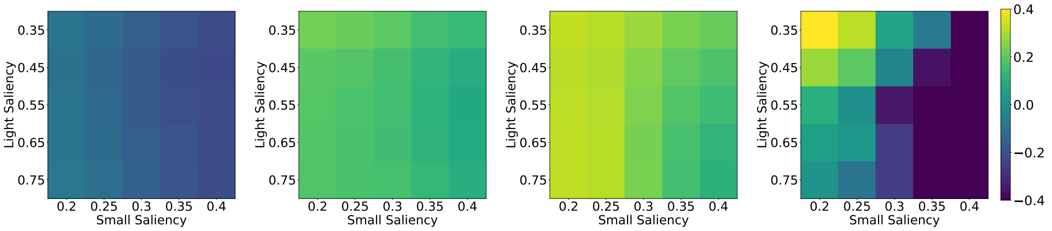
(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

Figure 16: Heatmap showing the difference in normalized probability of choosing "light" over "left," calculated as $\widehat{p} = \widehat{p}_{\text{light}} - \widehat{p}_{\text{left}}$. Darker colors indicate a preference for using the spatial term "left" over the size term "light."



(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

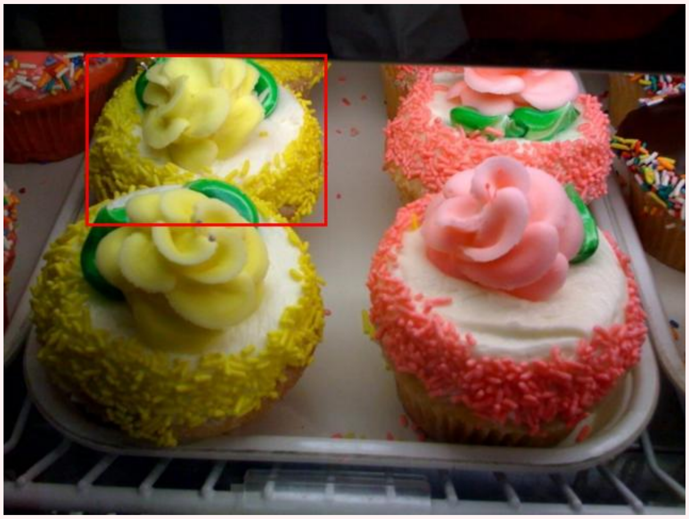
Figure 17: Heatmap showing the difference in normalized probability of choosing "light" over "square," calculated as $\widehat{p} = \widehat{p}_{\text{light}} - \widehat{p}_{\text{square}}$. Darker colors indicate a preference for using the spatial term "square" over the size term "light."



(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B. (d) Human.

Figure 18: Heatmap showing the difference in normalized probability of choosing "light" over "small," calculated as $\widehat{p} = \widehat{p}_{\text{light}} - \widehat{p}_{\text{small}}$. Darker colors indicate a preference for using the spatial term "small" over the size term "light."

Qualitative Comparison 1



Model Responses:

- **GPT-4o**: The object in the red box is a cupcake. It has yellow frosting shaped like a flower, with green leaves made of frosting on the side. The base of the cupcake is covered in yellow sprinkles.
- **LLaVA-7B**: The object in the red box is a cupcake with a yellow frosting and a green flower on top.
- **LLaVA-13B**: The object in the red box is a cupcake with a yellow frosting and a green rose on top. It is positioned in the top left corner of the box.
- **LLaVA-34B**: The object in the red box is a cupcake with a yellow frosting and a green ring around the top. It has a decorative flower made of frosting on top, and it's placed on a white tray. The cupcake is located in the middle of the image, slightly to the left of the center.
- **MiniCPM-V**: The object in the red box is a cupcake with yellow frosting, decorated with a green and white swirl on top. It has sprinkles around its base, giving it a festive appearance. This particular cupcake stands out due to its vibrant colors and unique decoration pattern compared to the others in the image.
- **XComposer**: The red-boxed object is a cupcake with yellow frosting and a yellow flower decoration.
- **CogVLM**: A yellow frosting flower.
- **GLaMM**: A small piece of a red candy.

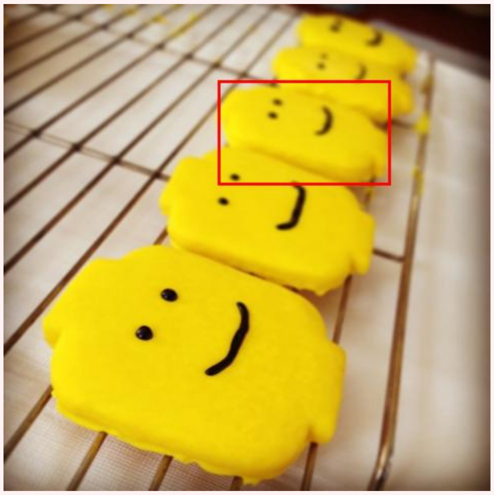
Human Written Responses:

- **User 1**: The farther yellow cake.
- **User 2**: The upper left yellow cake.
- **User 3**: The second yellow cupcake from the front.

Human Spoken Responses:

- **User 1**: The yellow cake at the upper left.
- **User 2**: The yellow cake in the top left corner.

Qualitative Comparison 2



Model Responses:

- **GPT-4o**: The object in the red box is a yellow, emoji-shaped cookie with a simple smiley face. It has two small black eyes, a smile, and is on a cooling rack. This cookie is positioned at the top of a stack of similar cookies, slightly offset to the left.
- **LLaVA-7B**: The object in the red box is a yellow cookie with a smiley face on it.
- **LLaVA-13B**: The object in the red box is a yellow cookie with a smiley face drawn on it. The smiley face consists of a black dot for the eye, a black line for the mouth, and a black dot for the nose. The cookie is placed on a cooling rack, which is a wire grid that allows air to circulate around the cookies to cool them evenly.
- **LLaVA-34B**: The object in the red box is a yellow cookie with a smiley face. It has two black dots for eyes and a black line for a mouth. The cookie is placed on a cooling rack, which is a common tool used in baking to allow air to circulate around the baked goods and prevent them from becoming soggy.
- **MiniCPM-V**: The object in the red box is a cookie with a yellow fondant covering and a smiley face design. The smile on this particular cookie is curved, giving it a cheerful expression. To locate this specific cookie, look for one that has a distinctively curved smile compared to the others, which have more rounded smiles.
- **XComposer**: The object in the red box is a yellow smiley face cookie that is placed on a wire cooling rack. This particular cookie is the fourth one from the left in the row of smiley face cookies. It has the typical features of a smiley face with two eyes, a nose, and a smiling mouth. The cookie appears to be of uniform shape and size, consistent with the other cookies in the row.
- **CogVLM**: Yellow cookie with black face.
- **GLaMM**: Sure, the segmentation result is [SEG].

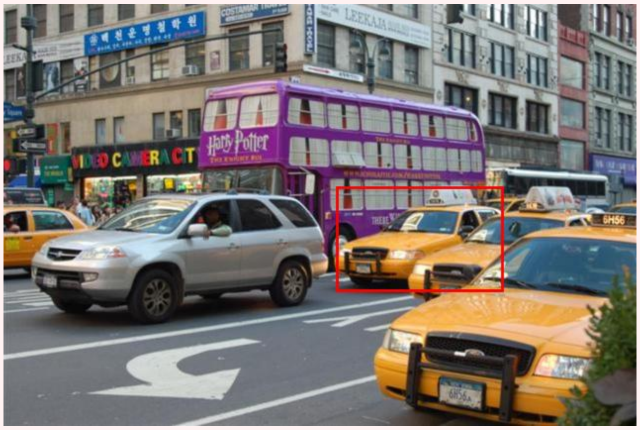
Human Written Responses:

- **User 1**: The third cake.
- **User 2**: The third biscuit from the bottom up.
- **User 3**: The middle smiling face.

Human Spoken Responses:

- **User 1**: The third biscuit from the bottom up.
- **User 2**: The third biscuit from the bottom up.

Qualitative Comparison 3



Model Responses:

- **GPT-4o**: The object inside the red box is a yellow New York City taxi cab. It is located in a busy street intersection, positioned between other vehicles. The taxi is facing towards the left, and part of the front side is visible within the red box. There is a recognizable “T” logo on the door of the cab, indicating it is a typical NYC taxi.
- **LLaVA-7B**: The red box is highlighting a purple double-decker bus with the words “Harry Potter” printed on its side. The bus is driving on a street with other vehicles and pedestrians.
- **LLaVA-13B**: The object inside the red box is a double-decker bus. It is purple in color and has the words “HARRY POTTER” written on its side. The bus is driving on a city street, and it is positioned behind a silver SUV.
- **LLaVA-34B**: The object inside the red box is a yellow taxi cab. It is positioned on the right side of the image, and it is one of several taxis visible in the scene. The taxi is a sedan with a distinctive yellow color, which is typical for taxis in many cities, and it has a visible license plate. The red box highlights this particular taxi among the other vehicles in the busy city street scene.
- **MiniCPM-V**: The object inside the red box is a yellow taxi with the license plate number ‘6H56A’. To locate this particular item, look for a yellow vehicle on the road that has the specified license plate. This taxi is part of New York City’s iconic fleet and can be identified by its distinctive color and design, which includes markings such as the medallion number on the side and possibly other city-specific signage or decals.
- **XComposer**: The object inside the red box is a yellow taxi cab. It can be identified by its distinctive color, which is a standard hue for taxis in New York City, and the number “BH56” visible on its side. The taxi is positioned in the middle of the street, amidst other vehicles, and is facing towards the right side of the image.
- **CogVLM**: A yellow taxi cab.
- **GLaMM**: Sure, the segmentation result is [SEG].

Human Written Responses:

- **User 1**: The yellow cab after that white car.
- **User 2**: The yellow car following the white car.
- **User 3**: The taxi nearest to the purple bus.

Human Spoken Responses:

- **User 1**: The second yellow car from the left.
- **User 2**: The yellow taxi just next to the purple bus.