FOCUS: Evaluating Pre-trained Vision-Language Models on Underspecification Reasoning

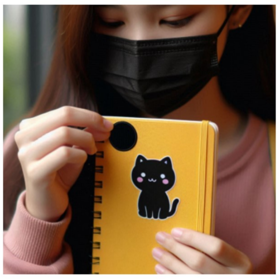
Kankan Zhou[1], **Eason Lai**[1], **Kyriakos Mouratidis**[1], **Jing Jiang**[1,2]
[1]School of Computing and Information Systems, Singapore Management University
[2]School of Computing, Australian National University
{kkzhou.2020, yibin.lai.2024, kyriakos, jingjiang}@smu.edu.sg, jing.jiang@anu.edu.au

Abstract

Humans possess a remarkable ability to interpret underspecified ambiguous statements by inferring their meanings from contexts such as visual inputs. This ability, however, may not be as developed in recent pre-trained vision-language models (VLMs). In this paper, we introduce a novel probing dataset called FOCUS to evaluate whether state-of-the-art VLMs have this ability. FOCUS consists of underspecified sentences paired with image contexts and carefully designed probing questions. Our experiments reveal that VLMs still fall short in handling underspecification even when visual inputs that can help resolve the ambiguities are available. To further support research in underspecification, FOCUS will be released for public use. We hope this dataset will inspire further research on the reasoning and contextual understanding capabilities of VLMs.

1 Introduction

Underspecification is common in human communication. It refers to the use of expressions that are intentionally left incomplete or vague. It relies on the listener's ability to infer the missing information from context, because humans are adept at interpreting underspecified communication by drawing on shared knowledge, prior experiences, and contextual cues such as body language, tone, or visual information. This capability enables efficient and flexible communication, where explicit details are often unnecessary for mutual understanding. For example, suppose two people are presented with the image in Figure 1 and one person mentions that "the girl attached the sticker to the notebook. It is yellow." The other person can easily infer that "it" here refers to the notebook instead of the sticker, based on the visual context. We use the term *underspecification reasoning* to refer to this ability to use external context, shared knowledge or logical reasoning to handle underspecified statements.



Textual Statement: The girl attached the sticker to the notebook. It is yellow.

Figure 1: An example of an underspecified sentence with a visual context from our FOCUS dataset.

While humans can easily handle underspecified sentences by incorporating visual context to fill in the missing information, do vision-language models (VLMs) have similar underspecification reasoning abilities? In an early work, Berzak et al. (2015) attempted to answer this question by building a specialized model that operates on first-order logic representations of different interpretations of an underspecified ambiguous sentence with the help of an external object detector. However, in recent years, general-purpose pre-trained VLMs such as LLaVA-NeXT (Liu et al., 2024) have demonstrated strong zero-shot vision-language understanding abilities such as visual question answering. A natural question to ask is whether these pre-trained VLMs can use visual input to disambiguate underspecified sentences. To the best of our knowledge, this research question has not been investigated.

In this paper, we empirically evaluate the capabilities of recent pre-trained VLMs w.r.t. underspecification, using a new probing dataset we introduce, called FOCUS (**F**ully **O**bserved **C**ontext with **U**nderspecified **S**entences). Our study is motivated by our belief that underspecification reasoning is a critical aspect of human communication and thus also a desirable ability for AI models.

Our FOCUS dataset consists of 2000 image-text pairs, where each pair contains an underspecified

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27565–27584
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

English sentence together with an AI-generated image that provides full visual context. FOCUS also comes with a set of probing questions. We use FOCUS to evaluate six pre-trained VLMs: GPT-4o-mini (OpenAI, 2024), LLaVA-NeXT (Liu et al., 2024), CogVLM (Wang et al., 2023), MoE-LLaVA (Lin et al., 2024), Qwen2-VL (Wang et al., 2024), and InstructBLIP (Dai et al., 2023). These models can directly interact with end-users through natural language conversations, making them readily available for wide adoption in end-user applications.

We center our evaluation on the following research questions. **R1:** Can a VLM correctly interpret an underspecified sentence when visual context is given? **R2:** How does a VLM's interpretation of an underspecified sentence without any visual context change after visual context is given? **R3:** Does an underspecified sentence affect a VLM's visual perception abilities?

Our experiments show that state-of-the-art pre-trained VLMs struggle to handle underspecified sentences effectively, even when provided with visual context. The best-performing models (GPT-4o-mini, LLaVA-NeXT, and CogVLM) manage to disambiguate less than 60% of the time. However, compared to cases where no visual context is available, we observe that several pre-trained VLMs still benefit from visual inputs when attempting disambiguation. On the other hand, underspecified sentences appear to negatively impact the models' visual perception abilities. In addition, we found that VLMs frequently provide inconsistent answers. Notably, all the underspecification cases we examine in this study involve ambiguous sentences. Although there are other forms of underspecification that do not involve ambiguity, those are beyond the scope of this work. FOCUS dataset will be made publicly available at https://github.com/K-Square-00/FOCUS.

2 Related Work

Underspecification. Semantic underspecification occurs when a sentence's meaning is incomplete, thus requiring contextual inference (Egg, 2010; Harris, 2020). Some recent studies have assessed pre-trained language models (LMs) on such scenarios. Wildenburg et al. (2024) introduced the DUST dataset and found that LMs struggle more with interpreting than identifying underspecified sentences. Similarly, Liu et al. (2023) studied the capabilities of LLMs to handle ambiguities through the task of textual entailment. However, both studies focus on language models rather than vision-language models. Pezzelle (2023) studied underspecification with VLMs and showed that SOTA VLMs struggle with underspecification. However, their study is limited to image-text association based on the CLIP model (Radford et al., 2021), without studying VLMs' disambiguation abilities. Our work, in comparison, evaluates SOTA VLMs' abilities to disambiguate underspecified ambiguous sentences with visual contexts.

It is worth noting that semantic underspecification is often studied alongside ambiguity. The key difference is that some underspecified sentences have a dominant interpretation (e.g., "Don't spend too much" (Wildenburg et al., 2024)), while ambiguous ones have multiple interpretations.

Pre-trained vision-language models. Early VLMs like ALBEF (Li et al., 2021), X^2-VLM (Zeng et al., 2023), and CLIP (Radford et al., 2021) lacked built-in visual question answering capabilities. With the rise of instruction-tuned LLMs, newer VLMs integrate visual perception via parameter-efficient fine-tuning of LLMs. Examples include BLIP-2 (Li et al., 2023a), LLaVA-Next (Liu et al., 2024), InstructBLIP (Dai et al., 2023), trainable visual experts like CogVLM (Wang et al., 2023), and multi-stage pipelines such as MoeLlaVA (Lin et al., 2024) and Qwen2-VL (Wang et al., 2024).

Evaluation of VLMs. VLMs have been evaluated in many different aspects such as compositional reasoning (Thrush et al., 2022), social bias (Zhou et al., 2022), and hallucination (Li et al., 2023b). However, evaluating VLMs' abilities to handle underspecification or ambiguity has not been systematically studied. Our work therefore investigates this underexplored problem.

3 The FOCUS Dataset

To systematically evaluate pre-trained VLMs' underspecification reasoning abilities using visual context, we first construct a new evaluation dataset. Previously, Berzak et al. (2015) released the LAVA dataset for similar purposes. However, the visual contexts in LAVA are videos, which cannot be handled by most current pre-trained VLMs. If we take still images from LAVA's videos, we encounter two problems: (1) Some of the ambiguous sentences in LAVA involve actions and require temporal information from the videos for disambiguation; and (2) the resolution of the still images extracted from LAVA's videos is low, making it often hard even for humans to rely on the visual information for disambiguation. We therefore need an appropriate new dataset.

In this section, we present FOCUS (Fully Observed Context with Underspecified Sentences), a dataset with 2000 (image, sentence) pairs to serve the purpose above. Sentences in FOCUS are underspecified with linguistic ambiguities, and the accompanying images show one of the two interpretations of the corresponding ambiguous sentence. FOCUS covers diverse types of linguistic ambiguities, ranging from prepositional phrase and verb phrase attachment to ellipsis. It makes a distinction between ambiguities that require visual context to resolve and ambiguities that can potentially be resolved through commonsense reasoning. The dataset also comes with a set of ternary (i.e., yes/no/unsure) questions to probe a VLM and test its ability to resolve the ambiguities in the underspecified sentences.

3.1 Underspecified Sentences

Following Berzak et al. (2015), we include the following six types of linguistic ambiguities that are commonly observed in underspecified sentences: prepositional phrase (PP) attachment, verb phrase (VP) attachment, conjunction, logical form, anaphora, and ellipsis. The details of each type can be found in Appendix A. Furthermore, we distinguish between two kinds of underspecification: We use the term *strong underspecification* to refer to ambiguities that require external context (in our case, visual context) to resolve. For instance, in the example of Figure 1, the pronoun "it" is ambiguous because it could refer to either the sticker or the notebook. This is an example of an anaphora ambiguity, but meanwhile, this ambiguity cannot be resolved until we can see whether it is the sticker or the notebook that is yellow. On the other hand, we use the term *weak underspecification* to refer to ambiguities that can likely be resolved using common sense and logical reasoning. The well-known Winograd Scheme Challenge (WSC) dataset (Bender, 2015) consists of such cases. For example, consider the sentence "the fish ate the worm because it was tasty". Although this sentence also contains anaphora ambiguity, by applying commonsense knowledge, one can almost be certain that the pronoun "it" here refers to the worm rather than the fish. As we can see, such cases of weak underspecification do not require additional visual input to help with the disambiguation. The reason we include examples of weak underspecification in FOCUS is that in case a pre-trained VLM has limited commonsense reasoning capabilities (which has been observed with some VLMs such as InstructBLIP and LLaVA (Zhou et al., 2023)), we wish to test whether the VLM can benefit from explicit visual context to handle weak underspecification.

Based on our classification above, we manually choose a subset of the sentence templates from LAVA (listed under the CC-by-4.0 license) to create our strong underspecification examples and a subset of the WSC sentences (listed under the CC-by-4.0 license) as templates to create our weak underspecification examples. When choosing sentence templates for both strong and weak underspecification, we cover the six different types of linguistic ambiguity. We then populate the templates with various common and realistic scenarios and objects to create a diverse dataset. Some examples of underspecified sentences in FOCUS are shown in Table 1.

3.2 Contextual Images

After collecting the set of underspecified sentences, our next step is to create accompanying images that correspond to one of the two interpretations for the strongly underspecified sentences, and to the more "reasonable" interpretation based on common sense for the weakly underspecified sentences. To facilitate image generation, we first expand each sentence into a fully specified version. We then input the fully specified sentences into ChatGPT, powered by DALL-E-3, to generate the corresponding images. In the case when a generated image does not correspond to the input sentence, we discard it and repeatedly refine the fully specified input sentence until a valid image is generated. On average, we have to discard about one invalid image for every 10 valid images generated. This image generation process is illustrated through an example in the Image Generation stage in Figure 2.

For a strongly underspecified sentence, we always generate two images corresponding to the two interpretations of the sentence. This is because previous studies have found that for some strongly underspecified sentences, humans may have preferred or default interpretations even when no visual cues are available (Dwivedi, 2013; AnderBois

Strong/Weak	Ambiguity Type	Example Underspecified Sentence
Strong Underspecification	Anaphora	The boy held the cup and the pen. It is white.
	Conjunction	The girl held the black bag and pen.
	Ellipsis	The girl is looking at the clock, also the man.
	Logical Form	The woman and the boy held a book.
	PP	The boy left the girl with tears.
	VP	The man looked at the girl walking on the street.
Weak Underspecification	Anaphora	The drain is clogged with hair. It has to be removed.
	Conjunction	The boy held the yellow hat and snow.
	Ellipsis	The woman is eating the cake, also the boy.
	Logical Form	The man and the woman wore a watch.
	PP	The boy answered the question with confidence.
	VP	The boy sat on the chair doing the homework.

Table 1: Examples of underspecified sentences.

et al., 2012). For example, for sentences like "every kid climbed a tree," it has been found that humans prefer a plural interpretation, i.e., interpreting the sentence as "every kid climbed a different tree." We used three examples of this kind to test VLMs and found that several VLMs showed a similar preference regardless of the image shown. (See Appendix B.) Therefore, to ensure that the contextual images are not biased towards any human- or VLM-preferred interpretations, we always include both interpretations of a strongly underspecified sentence when generating contextual images.

The image creation process is carried out independently by two authors of this paper, who cross-validate each other's selections. Only images agreed upon by both researchers to be valid images corresponding to the given sentences are retained. The Cohen's kappa score for the image creation process is 0.93, indicating a near perfect level of agreement.

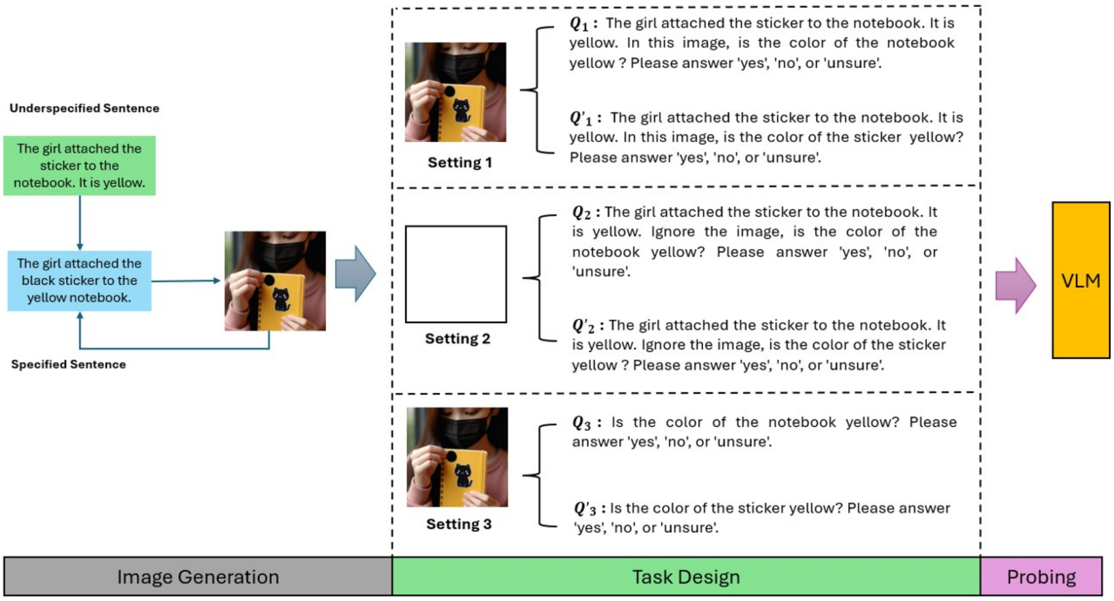
3.3 Probing Questions

Our primary research question is whether state-of-the-art pre-trained VLMs can effectively handle underspecified sentences when provided with visual context through an input image. Just like SOTA LLMs, these SOTA VLMs are typically used in a zero-shot manner through natural language prompts. Hence, we propose a set of natural language questions as probes to assess these VLMs' abilities to handle underspecification.

Pre-trained models like LLaVA-NEXT (Liu et al., 2024) often generate verbose responses to open-ended questions, which complicates the automated analysis of their answers. To address this issue, we design ternary questions that enable efficient response processing by checking if the answer starts with "yes," "no," or "unsure." The inclusion of "unsure" accounts for scenarios where ambiguity cannot be resolved and a human would answer "unsure."

The first set of questions is designed to directly check whether a pre-trained VLM can resolve the ambiguity in the context of the visual input, i.e., the accompanying image. For each sentence template, we design two question templates, each corresponding to one of the two interpretations of the ambiguous sentence. For example, given the ambiguous sentence "the woman and the boy held a book", one question asks whether the woman and the boy held the same book and the other asks whether the woman and the boy held different books. Examples of questions for each type of linguistic ambiguity can be found in Appendix C. We expect the model to answer "yes" to one of the questions and "no" to the other question, according to the image given. By measuring how often a VLM correctly answers these questions, we can judge the underspecification reasoning abilities of the VLM. We refer to this first setting as Setting 1.

To check whether a VLM indeed uses the provided visual context to facilitate disambiguation, we also want to test whether a VLM is confused and answers "unsure" when *no* visual context is given. Specifically, we expect a VLM to answer "unsure" when a *strongly* underspecified sentence is given. Therefore, we design another setting where the same questions are given to a VLM but an uninformative image such as a blank image in black or arguably a random image is provided. For weakly underspecified sentences, which can be disambiguated through commonsense reasoning, ideally a VLM should give a firm "yes" or "no" answer rather than "unsure" based on common sense, even when an uninformative image is provided. We refer to this setting where uninformative images are


Underspecified Sentence
The girl attached the sticker to the notebook. It is yellow.
The girl attached the black sticker to the yellow notebook.
Specified Sentence
Setting 1
Setting 2
Setting 3
VLM
Image Generation
Task Design
Probing

Q_1 : The girl attached the sticker to the notebook. It is yellow. In this image, is the color of the notebook yellow ? Please answer 'yes', 'no', or 'unsure'.

Q'_1 : The girl attached the sticker to the notebook. It is yellow. In this image, is the color of the sticker yellow? Please answer 'yes', 'no', or 'unsure'.

Q_2 : The girl attached the sticker to the notebook. It is yellow. Ignore the image, is the color of the notebook yellow? Please answer 'yes', 'no', or 'unsure'.

Q'_2 : The girl attached the sticker to the notebook. It is yellow. Ignore the image, is the color of the sticker yellow ? Please answer 'yes', 'no', or 'unsure'.

Q_3 : Is the color of the notebook yellow? Please answer 'yes', 'no', or 'unsure'.

Q'_3 : Is the color of the sticker yellow? Please answer 'yes', 'no', or 'unsure'.

Figure 2: Examples of probing questions.

provided as Setting 2.

Finally, we consider a third setting where we give the original informative image and directly ask the questions without giving the underspecified sentences. This makes sense for VLMs because the image contains all the information about the scenario described in the (underspecified) sentence. Therefore, even without the sentence, ideally a VLM should also be able to answer the question based on the visual input. We refer to this setting as Setting 3.

4 Measurements for Evaluation

We intend to use the images, underspecified sentences, and probing questions in FOCUS to evaluate pre-trained VLMs' underspecification reasoning abilities, or more concretely, to answer the research questions outlined in Section 1. In this section, we design several measurements that can facilitate our analysis later to answer those research questions. We first introduce some notation. Let (S, I) denote an underspecified sentence and its accompanying image that depicts one of the two interpretations of the sentence. Let Q_I denote the ternary question for (S, I) where an answer of "yes" matches the intended interpretation of S based on I. Let Q'_I denote the ternary question for (S, I) where an answer of "yes" corresponds to the alternative interpretation of S that is not depicted in I.

Answer consistency. In general, given a pair (S, I), we expect a VLM to give consistent answers to Q_I and Q'_I, that is, the answers to the two questions do not contradict each other. Concretely, we regard a VLM's answers to Q_I and Q'_I as *consistent* (but not necessarily correct) if and only if one of the following is true:

1) The answer to Q_I is "yes" and to Q'_I is "no."
2) The answer to Q_I is "no" and to Q'_I is "yes."
3) The answers to Q_I and Q'_I are both "unsure."

Any other answer combination is regarded as *inconsistent*.

R1: Can a VLM correctly interpret an underspecified sentence when visual context is given?

To answer this research question, we want to measure the percentage of probing questions that can be answered correctly and consistently based on the scenario depicted in the given image. We define the following metric:

Disambiguation Accuracy (DA): Given a model $\mathcal{M}$ and under Setting 1 defined in Section 3.3, the DA of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are consistent and the answers correspond to the interpretation of S depicted by I.

R2: How does a VLM's interpretation of an underspecified sentence change after visual context is given?

To answer this question, we want to observe model $\mathcal{M}$'s answers to the probing questions when the visual context is not given so that we can compare them with $\mathcal{M}$'s answers after the visual context is provided. When the visual context is not given, we expect $\mathcal{M}$ to be confused about the meaning of an underspecified sentence, at least for the strongly underspecified sentences. But this confusion may be manifested in different ways. Ideally we expect $\mathcal{M}$ to consistently answer "unsure" to both questions if it recognizes that the sentence is ambiguous, but if $\mathcal{M}$ gives inconsistent answers to the two questions, it can also be regarded as a sign of confusion. Furthermore, we also want to know how often $\mathcal{M}$ gives consistent answers that correspond to each interpretation of the sentence. In particular, this is useful when we examine the weakly underspecified sentences, because this allows us to check whether $\mathcal{M}$ uses any relevant prior commonsense knowledge to pick the more likely interpretation. We therefore define the following measurements:

Unsure Rate (UR): Given a model $\mathcal{M}$ and under Setting 2, the UR of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are both "unsure". Note that under Setting 2, instead of I, a blank image is given to model $\mathcal{M}$.

Inconsistency Rate (IR): Given a model $\mathcal{M}$ and under Setting 2, the IR of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are inconsistent (as defined earlier).

Consistency Rate 1 (CR1): Given a model $\mathcal{M}$ and under Setting 2, the CR1 of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are consistent and the answers correspond to the interpretation depicted by I.

Consistency Rate 2 (CR2): Given a model $\mathcal{M}$ and under Setting 2, the CR2 of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are consistent and the answers correspond to the alternative interpretation that is not depicted by I.

It is worth noting that for strongly underspecified sentences, when visual context is not given (i.e., under Setting 2), we expect $\mathcal{M}$ to have similar CR1 and CR2, because the model should not favor either interpretation in this case. For weakly underspecified sentences, on the other hand, $\mathcal{M}$ might favor the interpretation depicted by I even though I is not given, because I depicts the more "reasonable" interpretation based on common sense. So for weakly underspecified sentences, a model's CR1 might be higher than CR2.

R3: Does an underspecified sentence affect a VLM's visual perception abilities?

The probing questions in FOCUS can actually be answered based solely on the image given without the underspecified sentence. Therefore, the third research question asks whether giving an underspecified sentence to a model may affect the model's ability to use visual perception alone to answer the questions. To answer this research question, we can measure $\mathcal{M}$'s accuracy of correctly answering the questions based on the image *without* the underspecified sentence. Therefore, we introduce the following measurement:

Visual Perception Accuracy (VPA): Given a model $\mathcal{M}$ and under Setting 3, the VPA of $\mathcal{M}$ is the percentage of (S, I) pairs for which $\mathcal{M}$'s answers to Q_I and Q'_I are consistent and the answers correspond to the interpretation depicted by I. Note that under Setting 3, S is not given to the model.

5 Experiments

5.1 VLMs for Evaluation

We focus on six of the latest VLMs: GPT-4o-mini (OpenAI, 2024), LLaVA-NeXT (Liu et al., 2024), CogVLM (Wang et al., 2023), Qwen2-VL (Wang et al., 2024), MoE-LLaVA (Lin et al., 2024), and InstructBLIP (Dai et al., 2023). Model configuration details can be found in Appendix D. We also conduct a sanity check to verify the basic visual perception abilities of the VLMs and find that most models can achieve over 97% accuracy on object detection. See Appendix E for details.

5.2 Main Results

R1: Can a VLM correctly interpret an underspecified sentence when visual context is given?

The key metric here is DA. As can be seen from Table 2, across all models including GPT-4o-mini, when visual context is given, the performance measured by DA is considerably low, with scores ranging from 0.8% to 59.7%. For both strong and weak underspecifications, CogVLM, GPT-4o-mini, and LLaVA-NeXT are the stronger models, achieving DA scores between 50% and 60%. Meanwhile, Qwen2-VL, MoE-LLaVA, and InstructBLIP achieve much lower DA scores, all below 40%, with InstructBLIP giving extremely poor results of less than 2%. A closer look at InstructBLIP reveals that it suffers severely from generating inconsistent answers. These DA scores across all models

		With Visual Context	Without Visual Context			
Strong/Weak	**Model**	**DA**	**UR**	**IR**	**CR1**	**CR2**
Strong Underspecification	GPT-4o-mini	55.6	**41.2**	36.0	11.3	11.6
	LLaVA-NeXT	52.0	3.1	64.8	16.0	16.0
	CogVLM	**59.7**	0.0	58.3	20.8	20.8
	Qwen2-VL	29.0	25.9	66.0	3.5	4.5
	MoE-LLaVA	36.1	0.0	54.9	**22.8**	**22.2**
	InstructBLIP	0.8	0.0	**97.8**	1.1	1.1
Weak Underspecification	GPT-4o-mini	**59.1**	**22.2**	**28.4**	**49.1**	0.3
	LLaVA-NeXT	56.3	2.0	70.5	22.2	5.4
	CogVLM	56.5	0.0	41.8	47.4	**10.8**
	Qwen2-VL	39.8	9.9	67.0	18.8	4.3
	MoE-LLaVA	26.4	0.0	62.5	28.1	9.4
	InstructBLIP	1.4	0.0	**94.3**	3.7	2.0

Table 2: Comparison of VLMs' interpretations of underspecified sentences with and without visual context.

indicate that VLMs still struggle to interpret underspecified sentences when visual context is provided. Considering the models' high object detection accuracy, the results suggest that although current VLMs are adept at identifying visual elements in isolation, they fail to integrate these elements meaningfully with underspecified linguistic input.

R2: How does a VLM's interpretation of an underspecified sentence change after visual context is given?

To answer this question, we want to compare a VLM's interpretations of an underspecified sentence before and after visual context is given. We can use the defined metrics UR, IR, CR1 and CR2 to characterize a VLM's interpretation of a sentence without visual context. We separately discuss the strongly and the weakly underspecified sentences.

For strongly underspecified sentences, when no visual context is given, ideally we expect a VLM to consistently answer "unsure" (measured by the unsure rate UR), but inconsistent answers is also a (weaker) indication of unsureness (measured by the inconsistency rate IR). Therefore, we first examine the UR and IR scores of the models. we can see from Table 2 that GPT-4o-mini consistently answers "unsure" to 41.2% of the examples whereas the open-source models are much less likely to consistently answer "unsure". Meanwhile, GPT-4o-mini gives inconsistent answers to 36.0% of the examples whereas the open-source models give inconsistent answers to a large percentage of examples, ranging from 56.2% to 98.8%. These UR and IR scores suggest the following: (1) Neither GPT-4o-mini nor the open-source models will confidently pick one interpretation over the other for strongly underspecified sentences, which demonstrate that they have recognized the ambiguities in these sentences. (2) GPT-4o-mini has a much stronger ability to confidently answer "unsure", thus explicitly indicating the ambiguity detected, whereas the open-source models are generally not able to explicitly claim the ambiguity and can only imply the ambiguity by giving inconsistent answers. Next, we compare these models' behaviors with and without visual context. We can see that GPT-4o-mini, LLaVA-NeXT, CogVLM, and Qwen2-VL all have a higher DA score than the sum of CR1 and CR2 (which measures how often a model confidently picks one interpretation over the other without any visual context). This suggests that after visual context is given, these four models, particularly GPT-4o-mini and LLaVA-NeXT, have some abilities to use the visual context to resolve the ambiguities and pick the correct interpretation. For MoE-LLaVA, it has a weaker ability to detect ambiguity when no visual context is given (indicated by the highest CR1 and CR2 scores); therefore, given its low DA score, it is hard to conclude whether MoE-LLaVA is able to use visual context to resolve ambiguity. For InstructBLIP, its extremely low DA score shows that it cannot use visual context to resolve ambiguity.

For weakly specified sentences, because these sentences have a more "reasonable" interpretation, we compare the DA of a model when visual context is given with the CR1 of the model when visual context is not given. Recall that CR1 is the rate of the model interpreting the sentence based on common sense without visual context. We can see that for all models, CR1 is higher than CR2, suggesting that these models can use common sense to resolve ambiguity to some extent. For GPT-4o-

Underspecification	Model	DA	VPA
Strong Underspecification	GPT-4o-mini	55.6	48.9
	LLaVA-NeXT	52.0	63.6
	CogVLM	**59.7**	**68.5**
	Qwen2-VL	29.0	42.3
	MoE-LLaVA	36.1	47.5
	InstructBLIP	0.8	3.9
Weak Underspecification	GPT-4o-mini	**59.1**	40.1
	LLaVA-NeXT	56.3	51.4
	CogVLM	56.5	**54.8**
	Qwen2-VL	39.8	40.1
	MoE-LLaVA	26.4	38.4
	InstructBLIP	1.4	0.9

Table 3: Comparison of VLMs' abilities to answer visual questions with and without underspecified contexts.

mini, LLaVA-NeXT, CogVLM, and Qwen2-VL, their DA is clearly higher than CR1, indicating that the visual context helps them to further resolve the ambiguity.

Overall, for both strongly and weakly underspecified sentences, we conclude that GPT-4o-mini, LLaVA-NeXT, CogVLM, and Qwen2-VL can benefit from the provided visual context for underspecification reasoning, while MoE-LLaVA and InstructBLIP do not benefit from visual context.

R3: Does an underspecified sentence affect a VLM's visual perception abilities?

To answer R3, in Table 3 we compare DA with VPA, which measures the percentage of question pairs that can be consistently answered correctly given an image *without* the underspecified sentence as context. For strongly underspecified sentences, we can see that all models except GPT-4o-mini have a higher VPA score than DA score. This suggests that the strongly underspecified sentences have negatively affected these VLMs' abilities to use visual perception to answer questions, probably because the ambiguous context sentences have increased the cognitive load required to interpret the questions. For GPT-4o-mini, it is counter-intuitive to observe that without the strongly underspecified sentence as context, its ability to answer the questions based on the visual input has dropped. We do not have a good explanation for this and will leave further investigation as future work.

For weakly underspecified sentences, on the other hand, we find that LLaVA-NeXT, CogVLM, and Qwen2-VL have higher or similar DA scores compared with VPA scores. This suggests that the underspecified context sentences have not added much extra cognitive load for these VLMs to interpret the questions. In contrast, MoE-LLaVA has been negatively affected by the weakly underspecified context sentences, giving lower DA than VPA. Similar to the case with strongly underspecified sentences, GPT-4o-mini has a clearly higher DA than VPA for weakly underspecified sentences, which we do not have a good explanation for.

5.3 Failure Analysis

To analyze VLM errors with visual contexts, we examine results under Setting 1. Most failures (88.4%–99.8%) occur when models give inconsistent responses to question pairs, leading to misinterpretations of underspecified sentences.

Next, leaving out those cases of inconsistent responses, we look into all the cases that are incorrectly interpreted by CogVLM (the best performing model for strongly underspecified sentences). We find that interestingly 77% of these errors (which are 8.9% of total errors) are due to probing questions that ask whether two people are wearing (or carrying, holding, etc.) the same thing. An example is "the girl and the man wear a headphone. In this image, is the girl and the man wearing the same headphone?" Although the image shows the girl and the man each wearing their own headphones, the two headphones look the same. Because "the same headphone" can be interpreted as "same style or same model of headphone", VLMs may answer "yes", which is considered wrong based on our criterion but is arguably an acceptable answer due to the ambiguout meaning of "the same." The remaining 23% of non-inconsistency-related errors (2.7% of total errors) occur when VLMs are confused by complex, underspecified contexts (Figure 7). For instance, given "The boy looked at the girl drinking the milk.", most models infer the girl is drinking, contradicting the image. However, without the underspecified context, they correctly identify the boy as the drinker. This highlights how underspecified sentences can impair VLMs' visual perception. See Appendix F for details.

6 Conclusion

We have introduced the FOCUS dataset and a probing framework to assess VLMs on underspecification reasoning. Our results show that while visual input aids disambiguation, VLMs still struggle with underspecified sentences, even with visual context. Moreover, underspecification affects their visual perception. These findings reveal a gap in VLMs' ability to infer missing details, crucial for real-world communication. We hope FOCUS will drive research to improve multimodal reasoning and VLMs' comprehension of human communication.

7 Limitations

We acknowledge several limitations with the FOCUS dataset we have constructed and the evaluation we have conducted, and we discuss them in this section. One limitation of our study is the relatively small size of the FOCUS dataset compared to other larger-scale benchmarks. Because the dataset was designed to include highly unique and challenging underspecified scenarios, its creation was labor-intensive, thus limiting its size. While the small size of the dataset allowed for feasible manual annotation and ensured high data quality, we recognize the need for future expansion to increase its utility and diversity.

The FOCUS dataset is intended to be used for studying underspecification where the unspecified information can be inferred from the visual inputs. However, we acknowledge that when creating data points that require human involvement, particularly with visual content, social and cultural biases may be introduced to the data. In our work, the two annotators are Asian males. To reduce the possibility of introducing social and cultural biases, the annotators have chosen underspecified scenarios that are considered universal in different societies and cultures, based on their judgment. Nevertheless, we cannot guarantee that there is no hidden biases in the data, and this is a potential limitation of our work. Furthermore, by using ChatGPT to generate the contextual images, we have inevitably inherited any social or cultural biases that may exist within ChatGPT's image generation model (Cheong et al., 2024). This is another source of potential bias of our data that anyone using the dataset should be aware of.

Additionally, ambiguity in underspecified sentences is not always binary; there may be varying degrees to which something is considered underspecified. This makes it difficult to create data points that are unequivocally underspecified across different contexts. Furthermore, the trinary questions we designed to probe VLMs may not fully engage the models' reasoning capabilities. Some models may exhibit improved performance when prompted in a different manner, suggesting that the phrasing of queries can influence outcomes. Investigating more effective ways to prompt VLMs and exploring why they provide inconsistent responses to similarly phrased questions will be an important direction for future research.

Acknowledgements

This research was supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Proposal ID: T2EP20222-0047, Project ID: MOE-000440-00). Eason Lai and Kyriakos Mouratidis were supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award No. MOE-T2EP20121-0002). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the Ministry of Education, Singapore.

References

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2012. The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*.

David Bender. 2015. Establishing a human baseline for the winograd schema challenge. In *MAICS*, pages 39–45.

Yevgeni Berzak, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman. 2015. Do you see what I mean? visual resolution of linguistic ambiguities. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Marc Cheong, Ehsan Abedin, Marinus Ferreira, Ritsaart Reimann, Shalom Chalson, Pamela Robinson, Joanne Byrne, Leah Ruppanner, Mark Alfano, and Colin Klein. 2024. Investigating gender and racial biases in dall-e mini images. *ACM J. Responsib. Comput.*

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Veena D Dwivedi. 2013. Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS one*.

Markus Egg. 2010. Semantic underspecification. *Language and Linguistics Compass*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Zhang, Munan Ning, and Li Yuan. 2024. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sandro Pezzelle. 2023. Dealing with semantic underspecification in multimodal NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

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.

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.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Frank Wildenburg, Michael Hanna, and Sandro Pezzelle. 2024. Do pre-trained language models detect and understand semantic underspecification? ask the DUST! In *Findings of the Association for Computational Linguistics ACL 2024*.

Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. 2023. X^2-vlm: All-in-one pre-trained model for vision-language tasks. *Preprint*, arXiv:2211.12402.

Kankan Zhou, Eason Lai, and Jing Jiang. 2022. VL-StereoSet: A study of stereotypical bias in pre-trained vision-language models. 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 (Volume 1: Long Papers)*.

Kankan Zhou, Eason Lai, Wei Bin Au Yeong, Kyriakos Mouratidis, and Jing Jiang. 2023. ROME: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

A Common Types of Linguistic Ambiguity

Ambiguity Type	Description
Prepositional Phrase (PP) Attachment	Occurs when it is unclear which part of the sentence a prepositional phrase modifies. Example: "The boy looked at the woman with a telescope." – it is unclear whether "with a telescope" modifies "the boy" or "the woman."
Verb Phrase (VP) Attachment	Ambiguity arises regarding which verb phrase should be attached to a modifier. Example: The girl looked at the boy riding a bike." – it's ambiguous whether "riding a bike" modifies"the girl" or "the boy."
Conjunction	Happens when it is unclear how a conjunction relates to the items or clauses it connects. Example: "The girl held the blue pen and bag." – the ambiguity lies in whether the color blue applies to both pen and bag, or only to pen.
Logical Form	Ambiguity in how the logical structure of the sentence should be interpreted. Example: "The boy and the girl ate a cake." – this could mean the boy and the girl ate the same cake, or that each of them ate a different cake.
Anaphora	Refers to ambiguity concerning which antecedent a pronoun or a noun phrase refers to. Example: "The person held the box and the ball. It is white." – it is unclear whether "it" refers to the box or the ball.
Ellipsis	Occurs when the meaning is ambiguous due to omitted words or phrases. Example: "The dog is running toward the ball, also the boy." – it's unclear whether the boy is also running toward the ball, or the dog is also running toward the boy.

Table 4: Common types of linguistic ambiguity.

B Experiments for VLMs' default interpretation

To check if VLMs have a similar default interpretation of some strongly underspecified sentences like humans do, we borrow three examples of underspecified sentences from Dwivedi (2013). For each sentence, we generate two images, each corresponding to one of the two interpretations. We then design questions as shown in Figure 3. It has been shown by Dwivedi (2013) that when no visual cues are given, humans tend to interpret these sentences with the plural interpretation. VLMs' answers shown in Figure 3 suggest that some VLMs such as GPT-4o-mini and CogVLM also have similar preferences.

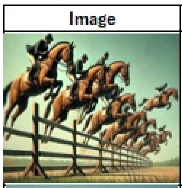
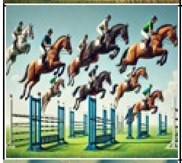
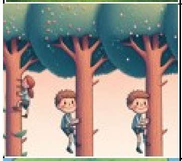
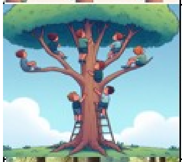
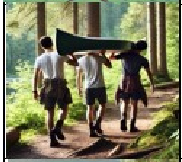
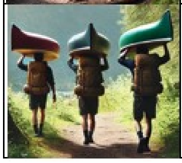
Image	Context	Questions	ChatGPT	CogVLM	InstructBLIP	LLaVA-NeXT	MoE-LLaVA	Qwen2-VL
	Every rider jumped a fence.	In this image, How many fences were jumped? Please answer 'one', or 'several'.	Several	Several	One	Several	One	Several
	Every rider jumped a fence.	In this image, How many fences were jumped? Please answer 'one', or 'several'.	Several	Several	One	Several	Several	Several
	Every kid climbed a tree.	In this image, How many trees were climbed? Please answer 'one' or 'several'.	Several	Several	One	One	Several	Several
	Every kid climbed a tree.	In this image, How many trees were climbed? Please answer 'one' or 'several'.	Several	Several	One	One	Several	Several
	Every camper carried a canoe.	In this image, How many canoes were carried? Please answer 'one' or 'several'.	Several	One	One	One	Several	One
	Every camper carried a canoe.	In this image, How many canoes were carried? Please answer 'one' or 'several'.	Several	Several	Several	Several	Several	Several

Figure 3: Experiment for VLMs' default interpretation.

C Sample questions with underspecified sentence and visual context in FOCUS

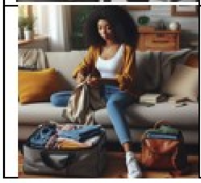
Image	Strong/Weak	Ambiguity Type	Underspecified Sentence	Disambiguation Accuracy (DA) Questions
	Weak Underspecification	Anaphora	The table won't fit through the doorway.	In this image, it is too wide. Does 'it' refer to the table? Please answer 'yes', 'no', or 'unsure'. In this image, it is too narrow. Does 'it' refer to the table? Please answer 'yes', 'no', or 'unsure'.
	Weak Underspecification	Conjunction	The boy held the black cup and tennis ball.	In this image, is the color of the tennis ball yellow? Please answer 'yes', 'no', or 'unsure'. In this image, is the color of the tennis ball not yellow? Please answer 'yes', 'no', or 'unsure'.
	Weak Underspecification	Ellipsis	The woman is riding the bike, also the girl.	In this image, are the woman and the girl riding the bike? Please answer 'yes', 'no', or 'unsure'. In this image, is the woman riding the bike and the girl ? Please answer 'yes', 'no', or 'unsure'.
	Weak Underspecification	Logical Form	The man and the woman wore a watch.	In this image, do the man and the woman wear different watches? Please answer 'yes', 'no', or 'unsure'. In this image, do the man and the woman wear the same watch? Please answer 'yes', 'no', or 'unsure'.
	Weak Underspecification	PP	The girl explained the problem with patience.	In this image, is the problem with patience? Please answer 'yes', 'no', or 'unsure'. In this image, is the girl with patience? Please answer 'yes', 'no', or 'unsure'.
	Weak Underspecification	VP	The woman sat on the sofa packing the bag.	In this image, is the sofa packing the bag? Please answer 'yes', 'no', or 'unsure'. In this image, is the woman packing the bag? Please answer 'yes', 'no', or 'unsure'.

Figure 4: Sample data for weak underspacification.

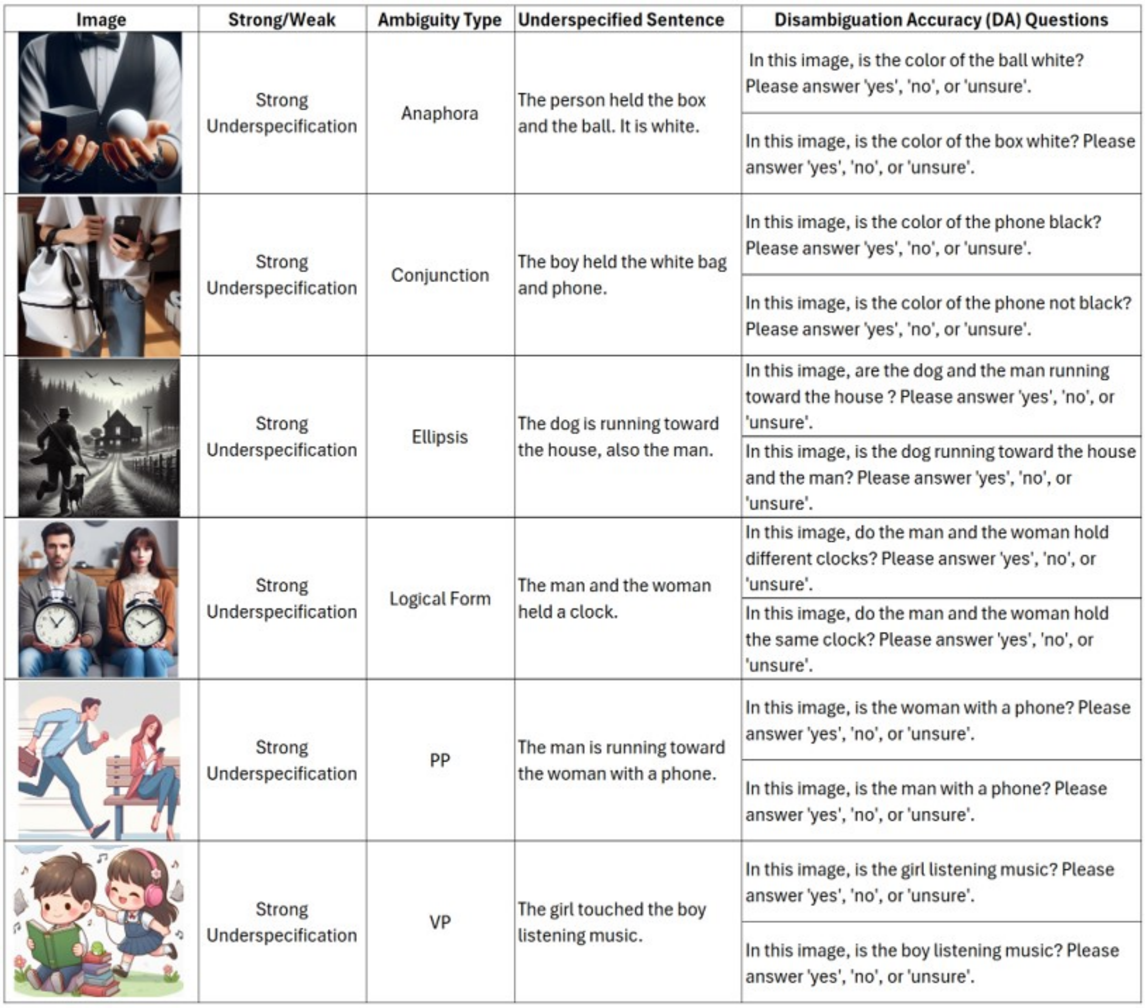


Image	Strong/Weak	Ambiguity Type	Underspecified Sentence	Disambiguation Accuracy (DA) Questions
	Strong Underspecification	Anaphora	The person held the box and the ball. It is white.	In this image, is the color of the ball white? Please answer 'yes', 'no', or 'unsure'. In this image, is the color of the box white? Please answer 'yes', 'no', or 'unsure'.
	Strong Underspecification	Conjunction	The boy held the white bag and phone.	In this image, is the color of the phone black? Please answer 'yes', 'no', or 'unsure'. In this image, is the color of the phone not black? Please answer 'yes', 'no', or 'unsure'.
	Strong Underspecification	Ellipsis	The dog is running toward the house, also the man.	In this image, are the dog and the man running toward the house ? Please answer 'yes', 'no', or 'unsure'. In this image, is the dog running toward the house and the man? Please answer 'yes', 'no', or 'unsure'.
	Strong Underspecification	Logical Form	The man and the woman held a clock.	In this image, do the man and the woman hold different clocks? Please answer 'yes', 'no', or 'unsure'. In this image, do the man and the woman hold the same clock? Please answer 'yes', 'no', or 'unsure'.
	Strong Underspecification	PP	The man is running toward the woman with a phone.	In this image, is the woman with a phone? Please answer 'yes', 'no', or 'unsure'. In this image, is the man with a phone? Please answer 'yes', 'no', or 'unsure'.
	Strong Underspecification	VP	The girl touched the boy listening music.	In this image, is the girl listening music? Please answer 'yes', 'no', or 'unsure'. In this image, is the boy listening music? Please answer 'yes', 'no', or 'unsure'.

Figure 5: Sample data for strong underspacification.

D Experiment Configuration

In this section, we provide additional details regarding the configuration of our computational experiments.

1. **Model Parameters and Computing Infrastructure**
   For open-source models, we conduct our experiments using the following variant of each model family.

Model	Model Size
cogvlm2	19B
MoeLlaVa	2.7B
LlaVaNext	7B
Qwen-VL	9.6B
InstructBlip	7B

   All experiments for open-source models have been conducted on two L40 GPUs, each equipped with 40 GB of memory. The runtime varies from 1 hour to 8 hours depending on the model.

   In addition to open-source models, we have conducted experiments using GPT-4o-mini, a proprietary model accessed via OpenAI's API. The exact model size of GPT-4o-mini is not publicly disclosed.

2. **Experimental Setup and Descriptive Statistics**
   To evaluate each model's performance under underspecified conditions using the FOCUS dataset, we utilized their default configurations. The detailed data pre-processing steps for the FOCUS testing dataset are described in Section 3.3. All results reported in this work are based on a single run.

3. **Packages Version**
   transformers=4.45.0 torch=2.5.1 cuda=11.8

Model	**Object Detection Accuracy**
GPT-4o-mini	98.9
LLaVA-NeXT	**99.0**
CogVLM	98.9
Qwen2-VL	97.3
MoE-LLaVA	97.2
InstructBLIP	74.6

Table 5: Object detection accuracy (scores in %).

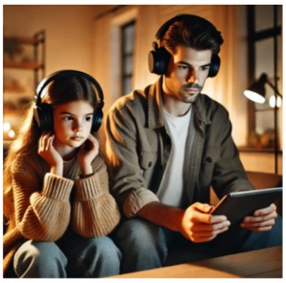
E Sanity Check for Basic Visual Perception

Sanity Check. Our experimental setup aims to determine whether VLMs can perform underspecification reasoning; however, this setup is based on the assumption that these VLMs have the basic visual perception capabilities. To test whether this assumption holds, we first perform a sanity check to see whether the pre-trained VLMs we choose have basic object detection capabilities. Specifically, we use yes/no questions to ask about the existence of the object mentioned in the probing question for all images of FOCUS. To make sure a model cannot "cheat" by answering "yes" to all questions, we also ask about the existence of objects that cannot be found in the images.

As shown in Table 5, most models demonstrate strong object detection capabilities, with scores above 95%, except for InstructBLIP, which achieves 74.6%. This indicates that the foundational object detection abilities of these models are generally robust. However, the underperformance of InstructBLIP, particularly with a score of 75.4% for detecting non-existent objects, highlights a tendency towards over-recognition or "hallucination," where the model incorrectly identifies objects that are not present in the image.

F Examples of Errors

In this appendix we show two examples of errors made by VLMs. The first example shown in Figure 6 is an example where the meaning of “the same” in the probing question is ambiguous. It can be interpreted as “a common headphone” or “the same model of headphones”. The second example shown in Figure 7 is an example where some of the VLMs can correctly answer the question when the underspecified sentence is not provided as the context, but when the sentence is provided, the VMLs become confused and give the wrong answer. The example shows how underspecified contextual sentences might hurt the visual perception of VLMs.



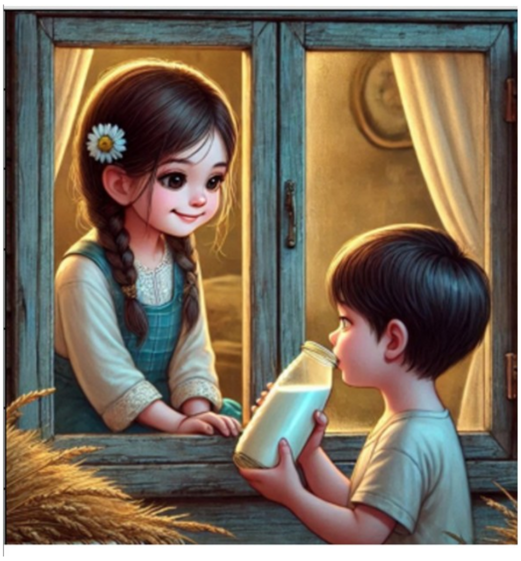
[Question] The girl and the man wore a headphone. In this image, Do the girl and the man wear the same headphone ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	Yes	**LLaVA-NeXT :**	No	**InstructBLIP:**	yes
Qwen2-VL :	Yes	**GPT-4o-mini:**	Unsure	**CogVLM:**	Yes

[Question] The girl and the man wore a headphone. In this image, Do the girl and the man wear different headphones ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	No	**LLaVA-NeXT :**	No	**InstructBLIP:**	yes
Qwen2-VL :	Yes	**GPT-4o-mini:**	Yes	**CogVLM:**	No

Figure 6: Sample VLMs Responses 1.



[DA Question] The boy looked at the girl drinking the milk. In this image, Is the girl drinking the milk ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	Yes	**LLaVA-NeXT :**	Yes	**InstructBLIP:**	Yes
Qwen2-VL :	Yes	**GPT-4o-mini:**	No	**CogVLM:**	Yes

[VPA Question] Is the girl drinking the milk ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	No	**LLaVA-NeXT :**	No	**InstructBLIP:**	Yes
Qwen2-VL :	Yes	**GPT-4o-mini:**	No	**CogVLM:**	No

[DA Question] The boy looked at the girl drinking the milk. In this image, Is the boy drinking the milk ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	No	**LLaVA-NeXT :**	No	**InstructBLIP:**	Yes
Qwen2-VL :	No	**GPT-4o-mini:**	No	**CogVLM:**	No

[VPA Question] Is the boy drinking the milk ? Please answer 'yes', 'no', or 'unsure'.

MoE-LLaVA:	Yes	**LLaVA-NeXT :**	Yes	**InstructBLIP:**	Yes
Qwen2-VL :	Yes	**GPT-4o-mini:**	No	**CogVLM:**	Yes

Figure 7: Sample VLMs Responses 2.