This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VAGUE: Visual Contexts Clarify Ambiguous Expressions

Heejeong Nam*[1], Jinwoo Ahn*[2], Keummin Ka[3], Jiwan Chung[3], and Youngjae Yu†[3]

[1]Brown University
[2]UC Berkeley
[3]Yonsei University

Abstract

Human communication often relies on visual cues to resolve ambiguity. While humans can intuitively integrate these cues, AI systems often find it challenging to engage in sophisticated multimodal reasoning. We introduce VAGUE, a benchmark evaluating multimodal AI systems' ability to integrate visual context for intent disambiguation. VAGUE consists of 1.6K ambiguous textual expressions, each paired with an image and multiple-choice interpretations, where the correct answer is only apparent with visual context. The dataset spans both staged, complex (Visual Commonsense Reasoning) and natural, personal (Ego4D) scenes, ensuring diversity. Our experiments reveal that existing multimodal AI models struggle to infer the speaker's true intent. While performance consistently improves from the introduction of more visual cues, the overall accuracy remains far below human performance, highlighting a critical gap in multimodal reasoning. Analysis of failure cases demonstrates that current models fail to distinguish true intent from superficial correlations in the visual scene, indicating that they perceive images but do not effectively reason with them. We release our code and data at https://hazel-heejeong-nam.github.io/vague/.

1. Introduction

Human communication is inherently contextual; for example, exclaiming "Hey, this is a disaster!" upon seeing a cluttered room conveys frustration or exaggeration rather than referring to an actual catastrophe. Without surrounding cues, textual dialogues can be *ambiguous*, making it difficult for models to accurately capture intent and nuance.

We consider the case of *visual* contextual cues. Consider Fig. 1, which depicts a speaker making a remark in a certain situation. Without specifying contexts introduced from visual cues, the speaker's intention can vary, thus remaining ambiguous. This implies that the visual contexts play important roles in communication, raising the question: can AI systems integrate visual cues with ambiguous dialogue to infer the speaker's intent?

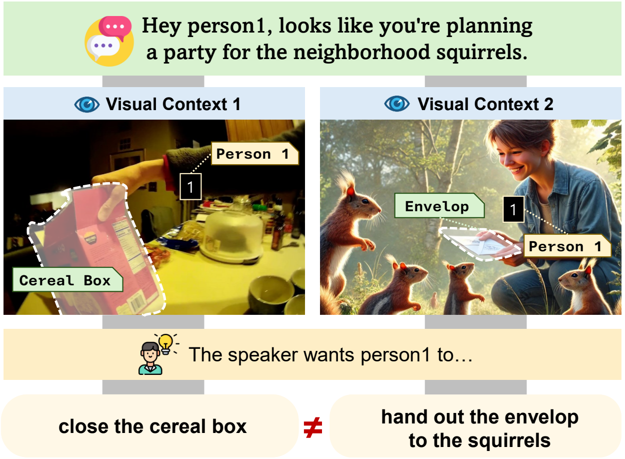

Hey person1, looks like you're planning a party for the neighborhood squirrels.
Visual Context 1
Visual Context 2
Person 1
1
Cereal Box
Envelop
1
Person 1
The speaker wants person1 to...
close the cereal box
≠
hand out the envelop to the squirrels

Figure 1. A motivating example demonstrating the importance of visual context in understanding intention. Without a predetermined context, a single expression can convey multiple different intentions. The textual expression and Context 1 are from our dataset, while Context 2 is generated using DALL·E 3 [2] to help understanding.

We introduce *Visual Contexts ClArify ambiGUous Expressions* (VAGUE), a benchmark consisting of 1.6K ambiguous textual expressions, each paired with a single image. VAGUE aims to model diverse and natural human-to-human interactions by setting each image as the speaker's viewpoint, where the speaker implicitly requests a certain action from a person within their field of view. We define the problem addressed through this setup as Multimodal Intention Disambiguation (MID), which involves reasoning about the most plausible request conditioned on visual context. Each sample in VAGUE is annotated with four multiple-choice candi-

*These authors contributed equally.
†Corresponding author.

dates, ensuring clarity and preventing multiple valid answers caused by paraphrasing or hierarchical inclusion of meaning. The dataset is meticulously curated to ensure visual dependency; the ground-truth candidate is only preferable when considering the visual context. VAGUE includes visual scenes from both artificial sources (Visual Commonsense Reasoning [46]) and real-world scenarios (Ego4D [41]), capturing a broad spectrum of scene complexity and naturalness. The textual expressions in VAGUE are initially generated by GPT-4o [31] following instructions, then reviewed through extensive human rating and filtering to ensure naturalness and alignment with the corresponding images.

Experiments on VAGUE demonstrate that existing multimodal AI models struggle to infer a speaker's true intent in a multimodal setting. First, although models can leverage visual context—as seen by a performance progression from text-only Language Models (LMs) to pipelined Socratic Models (SMs) [47] and ultimately to end-to-end Visual Language Models (VLMs)—their overall accuracy remains significantly lower than that of humans, indicating a failure to capture the true intent. A closer analysis of failure cases reveals that the primary source of error is the models' inability to distinguish the true intent from a superficial understanding of the visual context. In other words, even though these multimodal systems can perceive the image content, they cannot effectively use this information to reason about the speaker's true intent.

In conclusion, we introduce a benchmark that exposes the limitations of current models in integrating visual cues with intent comprehension and identifies their primary failure mode. We anticipate that VAGUE will serve as a testing ground for the development of future multimodal conversational or embodied agents—systems that combine robust visual perception with nuanced conversational reasoning to effectively respond to user requests in complex scenes.

Our contributions are threefold:

- VAGUE: a novel benchmark for evaluating multimodal intention disambiguation. Validated through extensive human filtering, VAGUE is designed for robust quantitative assessment by ensuring both the ambiguity of queries and the visual (in)dependency to answer candidates.
- Carefully curated 1,677 scene images sourced from VCR [46] and Ego4D [41], capturing a wide range of scene complexity, diversity, and naturalness to ensure VAGUE's generalizability across various contexts.
- Experimental results highlighting a critical challenge in multimodal intention disambiguation: while existing models can perceive visual cues, they fail to effectively integrate this information into reasoning to deduce the speaker's true intent.

2. Related Work

2.1. Multimodal Theory of Mind

Theory of Mind (ToM) refers to the ability to infer and reason about the intentions of others based on available information [34], where recent language models still struggle with relevant tasks [7] highlighting the need for dedicated research in this area. Initially, various methods and benchmarks have been proposed in unimodal settings, relying on text-based approaches [11, 36]. However, these methods often fail to capture the richness of real-world interactions, which often require integrating both linguistic and visual cues.

Moving beyond text-only contexts, recent work has incorporated visual information. MMToM [17] introduces a benchmark where models must process both visual and textual cues to solve question-answering tasks related to ToM. The BOSS dataset [9] is a multimodal dataset collected in situations where nonverbal communication is required. It is used to evaluate whether human beliefs can be inferred based on nonverbal cues during social interactions. Similarly, Chen et al. (2024) [5] propose a Video ToM model that leverages key video frames and transcripts, demonstrating improved reasoning on the Social-IQ 2.0 dataset [45]. MuMA-ToM [37] further extends this direction by assessing ToM reasoning in multi-agent interactions, evaluating a model's ability to infer human beliefs and goals based on video and text inputs. MToMnet [3] introduces a ToM-based neural network that integrates contextual cues, such as scene videos and object locations, with person-specific cues, to predict human beliefs in specific scenarios.

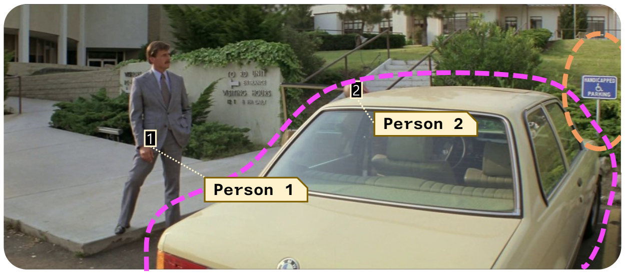
However, progress in multimodal ToM remains constrained not only by the scarcity of high-quality datasets [5] but also by the lack of explicit consideration for the ambiguity and indirectness inherent in human communication.

2.2. Multimodal Implicature Understanding

Implicature and the ambiguity that arises from it naturally emerge in everyday human conversation, requiring pragmatic understanding [38]. Early research on implicature understanding has primarily been conducted in text-only settings [27, 30, 40], with some studies specifically focusing on figurative language and metaphor [4, 21, 39]. However, since the ambiguity of standalone text is inherently limited, recent studies have extended to multiple modalities. One example is multimodal sarcasm understanding (MSU). WITS [19] and MOSES [20] are benchmarks for sarcasm explanation, both providing the speaker's emotion and voice tone as cues. DocMSU [8] is a document-level benchmark for sarcasm localization and detection. To improve MSU, EDGE [32], a graph-based approach, achieved strong performance. UR-FUNNY [13] is a benchmark for multimodal humor comprehension, incorporating facial expressions and voice tones as in MSU [19, 20]. Hessel et al. (2023) [14] intro-

Q. Select the option that best explains the **underlying intention** of the utterance based on the given image.

Hey person1, spot the difference, this parking's a bit too special isn't it?



a) The speaker wants person1 to admire the unusually decorated motorcycle in the parking lot.

b) The speaker wants person1 to enjoy playing a puzzle game and spot the difference.

✓ c) The speaker wants person1 to move the sedan because it's in a handicapped parking spot.

d) The speaker wants person1 to move the sedan because it's parked in front of a fire hydrant.

Figure 2. Description of the Multimodal Intention Disambiguation (MID) task in a Multiple-Choice Question format: Given an input image (I) and an indirect expression (p_i), the goal is to infer the speaker's hidden intent (T) and select the most likely answer.

duced a benchmark derived from a Cartoon Caption Contest, exploring humor identification and explanation. Baluja et al. (2024) [1] demonstrated that models benefit from multimodal cues in humor understanding. Memes also involve implicature, with multimodal datasets such as MemeCap [15] and MultiBully-Ex [16].

However, the cues used in multimodal implicature understanding remain simple, primarily appearing in images with a single main object or person, overlooking the importance of interactions between multiple objects and people in real-world scenarios. These limitations underscore the need for more complex cues, as addressed in VAGUE.

3. Multimodal Intention Disambiguation

In this section, we outline the structure and rationale behind the format of our primary task, which we term Multimodal Intention Disambiguation (MID). Then, we further specify the necessary components that form the basis of the task.

3.1. Problem Setting

Each MID problem comprises an input image I, a direct text expression p_d, and an indirect text expression p_i. Here, the direct expression p_d clearly shows the underlying intention of the corresponding p_i and serves as an essential intermediate step of generating p_i.

To clarify our problem, we assume that all reasoning is confined to the depicted scene and that each expression is spoken by a human who intends for the listener to take a particular action based on the situation. The ultimate objective of the task is to interpret the hidden intention T effectively by leveraging the contextual cues within the image.

To evaluate how well models capture such intentions, we adopt a multiple-choice (MCQ) format as the primary setup, as shown in Fig. 2. This decision reflects the fact that certain prompts can lead to multiple plausible outcomes, driven by hierarchical relations (e.g., pick up the *chips - snack - food*) or by the inherent uncertainty of what action best satisfies the speaker's goal (e.g., an indirect prompt complaining about darkness could be addressed by either turning on a light or opening curtains). Exploring all possible valid interpretations is labor-intensive and often infeasible. Consequently, each MID instance is presented as four distinct options, one correct and three intentionally designed to be incorrect for different reasons (see Sec. 4.2.3), challenging models in both linguistic and visual reasoning.

Formally, let C be the set of all multiple-choice options c_n. Given an image I and an indirect prompt p_i, the task is to select the most valid interpretation of p_i from the predefined options, conditioned on the visual context in I. We define this task as follows:

$$T(I, p_i) := \text{argmax}_{c_n \in C} \Pr(c_n \mid I, p_i). \tag{1}$$

3.2. Direct and Indirect Expressions

By the design of our task, curating effective input prompts p_d and p_i is crucial for ensuring accurate interpretation. What makes a *good* prompt, though? In this section, we define and explain the criteria that both direct and indirect expressions must satisfy. For more details on good and bad examples for each criterion, please refer to our Appendix A.

3.2.1. Directness: Relevance and Solvability

Relevance The direct prompt is an utterance from the speaker that explicitly conveys its intended meaning without ambiguity. However, it is equally important that this intention aligns with the visual context of the scene. For example, a direct prompt p_d such as "Hey person1, I want you to stop the fireworks" clearly expresses its intended action. However, if the corresponding image, as shown in Fig. 2, contains no elements related to fireworks, the prompt is misaligned with the scene. Thus, a direct prompt must not only reveal its intention but also maintain relevance to the image. In the context of our task, **relevance** is determined by whether a human can reasonably establish a connection between the prompt and the depicted scene.

Solvability Relevance alone does not guarantee that a prompt is useful. As outlined in Sec. 3.1, the prompt p_d must explicitly request an action that the listener can reasonably perform. This introduces **solvability**, which requires that the prompt present a clear and actionable problem. A solvable prompt defines a specific issue that can be addressed independently, ensuring that the listener is not left with multiple competing actions to choose from.

3.2.2. Indirectness: Consistency and Ambiguity

Consistency Indirect prompts are designed to obscure their true intention, but they must still convey the same underlying intention as their direct counterpart—essentially requesting the same solution. Since indirect prompts are derived from direct prompts, consistency serves as a key criterion. We define a direct prompt p_d and an indirect prompt p_i as consistent if their interpretations could *potentially* align in intention. The term "potentially" is used because indirect prompts, by nature, may have multiple valid interpretations. For instance, in the earlier "this is a disaster!" example, it conveys distress but allows for multiple reasonable responses, such as cleaning the room or providing reassurance.

Ambiguity If an indirect prompt is entirely consistent with its direct counterpart without introducing any additional complexity, it becomes indistinguishable from a direct prompt. Therefore, an indirect prompt should conceal its underlying intention, which we define as ambiguity. The key principle behind this criterion is that neither the specific action required nor the key entity involved should be explicitly or implicitly mentioned within the prompt. Once these two elements are concealed, further refinements—such as adjusting the tone to be more indirect, sarcastic, or humorous—can enhance the overall nuance and difficulty of interpretation.

4. VAGUE Benchmark Construction

VAGUE is a novel benchmark that extends single-modal or simple multimodal ambiguity to more realistic domains and evaluates whether concurrent vision-language models can perform human-like reasoning with complex visual contexts. It comprises 1,677 images, with 1,144 sourced from the VCR [46] dataset and 533 from Ego4D [41], covering diverse contextual scenarios as well as real-world human interactions. On average, VAGUE contains seven objects and four people per image. Each image is paired with a direct expression p_d, an indirect expression p_i, and four multiple-choice answers, along with relevant meta-information. All textual components are generated using GPT-4o, then refined through extensive human rating, selection, and filtering, ensuring a carefully curated benchmark dataset for testing and advancing multimodal reasoning. We provide detailed benchmark statistics and a diversity analysis in Appendix B.2.

4.1. Visual Data Curation

4.1.1. Sampling

VCR [46] The VCR dataset consists of 110K movie scenes sourced from the Large Scale Movie Description Challenge [35] and YouTube clips. These images are curated based on an "interestingness" criterion [46], ensuring the presence of at least two people, which promotes interactive scenarios. To prevent redundancy in our dataset, we sample 10K images while carefully avoiding neighboring frames, as adjacent frames exhibit minimal variation. This selection process preserves the contextual diversity of the dataset while maintaining its focus on complex, multi-entity interactions.

Ego4D [41] While VCR provides a wide range of contextual diversity, it often includes artificially composed settings that may not fully capture real-world interactions. To address this, we integrate frames from the Ego4D dataset, which offers a more naturalistic depiction of human interactions. We specifically leverage the AV (Audio-Visual), which indicate conversational exchanges between individuals, to ensure the presence of people in the selected frames. Similar to VCR, we avoid neighboring frames to maintain diversity and filter out heavily blurry images to enhance data quality. This process results in 888 candidate images from 94 videos, which serve as the basis for further text processing.

4.1.2. Object Extraction

To ensure the complexity of the visual information, we extract a list of physical objects present in each image using a tagging model, RAM [49]. This step allows us to easily identify scenes with sufficient visual detail. In the case of VCR, many scenes are relatively simple, often containing only a few objects. Therefore, we sort VCR images by the number of detected objects and retained the top 4,000 as candidates for text processing, ensuring that our benchmark primarily consists of rich visual cues in contextually diverse scenes. Please refer to Appendix B.3 for more details.

4.1.3. Person Indicator

In our task, we assume that the speaker is outside the scene, viewing the image and talking to a person in it. However, identifying the addressee is not always straightforward, as images mostly contain multiple individuals. While grounding the specific person referenced in the utterance could introduce additional complexity, it is not the primary focus of our evaluation. To clarify the listener, we assign an indicator tag to each person in the image as shown in Fig. 3. For VCR, we use their existing annotations [46], while for Ego4D, where bounding boxes are not fully available, we employ YOLOv11 [18] to detect and annotate humans. While this provides a straightforward method for grounding the target person, it requires models to perform basic Optical Character Recognition (OCR). Therefore, we conduct exper-

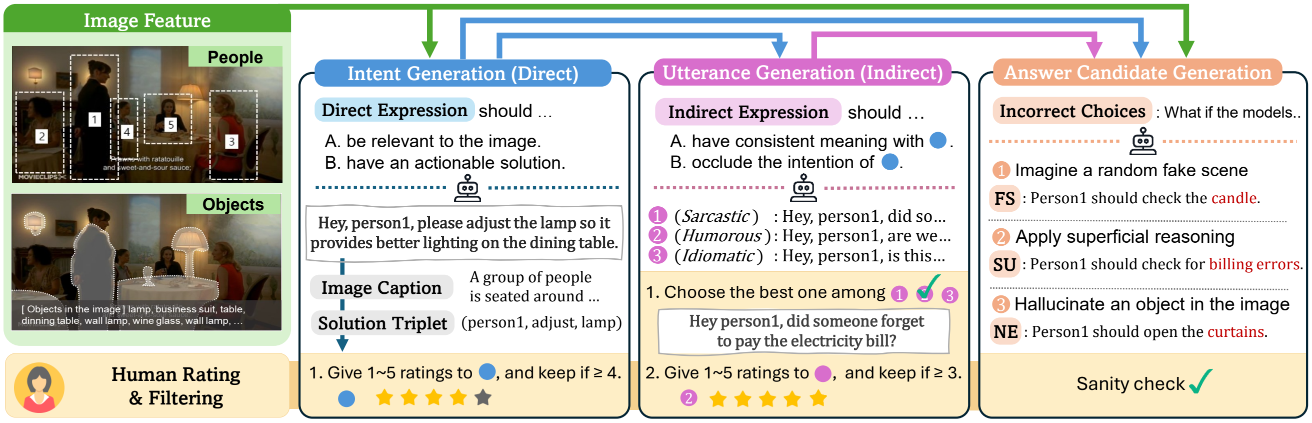


Image Feature
People
1
2
3
4
5
Prawns with ratatouille and sweet-and-sour sauce;
MOVIECLIPS
Objects
[Objects in the image] lamp, business suit, table, dinning table, wall lamp, wine glass, wall lamp, …
Human Rating & Filtering
Intent Generation (Direct)
Direct Expression should …
A. be relevant to the image.
B. have an actionable solution.
Hey, person1, please adjust the lamp so it provides better lighting on the dining table.
Image Caption
A group of people is seated around …
Solution Triplet
(person1, adjust, lamp)
1. Give 1~5 ratings to ●, and keep if ≥ 4.
Utterance Generation (Indirect)
Indirect Expression should …
A. have consistent meaning with ●.
B. occlude the intention of ●.
1 (Sarcastic) : Hey, person1, did so…
2 (Humorous) : Hey, person1, are we…
3 (Idiomatic) : Hey, person1, is this…
1. Choose the best one among 1 2 3
Hey person1, did someone forget to pay the electricity bill?
2. Give 1~5 ratings to ●, and keep if ≥ 3.
2
Answer Candidate Generation
Incorrect Choices : What if the models..
1 Imagine a random fake scene
FS : Person1 should check the candle.
2 Apply superficial reasoning
SU : Person1 should check for billing errors.
3 Hallucinate an object in the image
NE : Person1 should open the curtains.
Sanity check ✓

Figure 3. Overview of the data generation process. Based on human-defined criteria and instructions, GPT-4o [31] generates initial data, which are then rated and filtered by humans to ensure quality. Since generating high-quality indirect expressions from raw images is challenging, the process follows these steps: generating a direct expression and intention from the image, creating an indirect expression using information from the previous step, and producing answer candidates based on all the information gathered so far. In the answer candidates, FS stands for Fake Scene Understanding, SU stands for Superficial Understanding, and NE stands for Nonexistent Entity. FS evaluates global hallucination, where the model misinterprets visual context from an entirely different image (e.g., *an outdoor camping scene with a candle*), while NE addresses local hallucination, where only a specific object is replaced with a fabricated one (e.g., *curtains*).

iments to evaluate the OCR capabilities of the models used in our study. See Appendix B.4 for details.

4.2. Multimodal Expression Synthesis

As shown in Fig. 3, VAGUE's textual expressions are generated first by the model and undergo an extensive process of human rating and filtering. We use GPT-4o [31] for all text processing. The instructions used for generating direct (p_d) and indirect (p_i) expressions are provided in Appendix B.5, while the instructions for generating answer candidates for multiple-choice questions are detailed in Appendix B.7.

4.2.1. Direct Expressions

The direct expression p_d serves as a crucial foundation for crafting the indirect one p_i, as both share the same underlying intention. To ensure that p_d adhere to the principles of relevance and solvability discussed in Sec. 3.2.1, we generate p_d conditioned on the input image I and a task prompt that explicitly defines these criteria.During the generation process, we also instruct the model to output a solution triplet in the format: (subject, action, object). Since direct expressions explicitly state their intentions, extracting each component of the solution triplet is straightforward. To maintain consistency with the visual context, the ”object” in the triplet is restricted to physical objects we extracted in Sec. 4.1.2.

After generating p_d for all candidate images, human raters evaluate each prompt based on relevance and solvability, assigning scores from 1 to 5. Only those prompts that receive a rating of 4 or 5 are retained for further use. The detailed rating criteria for human verification and an example of the rating process are provided in Fig. J13.

4.2.2. Indirect Expressions

To ensure that the indirect expressions p_i maintain both ambiguity and fluency while aligning with the true intent T, we adopt a two-stage process: **proposal** and **selection**.

In the initial step, the model is prompted to generate three distinct candidate options. Each candidate follows the criteria outlined in Section 3.2.2, but with explicit instructions to incorporate different linguistic strategies: sarcasm, humor, and meme/idiomatic expressions, respectively. This approach ensures diversity in the generated responses. In the second step, human annotators evaluate the three candidates and select the one that best aligns with the intended indirectness. The selected prompt is then rated on a scale from 1 to 5 based on more specific criteria. Only those prompts that receive a score of 3 or higher are retained for use in the dataset. The detailed rating criteria for human verification and an example of rating process are provided in Fig. J14

4.2.3. Counterfactual Choices

Generating high-quality counterfactual choices is crucial. To enable detailed analysis of model weaknesses in multimodal intent disambiguation, we design interpretable counterfactual choices that provide more plausible alternatives.

Fake Scene Understanding The first counterfactual choice is an interpretation that could arise when the model largely misinterprets the image. This process is conducted in two steps. In the first step, a fake caption is generated by assuming an imaginary scene that can be aligned with the indirect expression but is inconsistent with the true intent. The caption of fake scene is then combined with the speaker's indirect statement to derive the most likely interpretation.

Superficial Understanding The subsequent choice corresponds to an interpretation generated when the model fails to deeply reason about the implicit intent of the sentence and instead relies on surface-level meaning. We enforce the model to focus only on the literal wording, without considering any implied or deeper meaning of the indirect sentence. These answer choices are generated alongside the indirect expression p_i. During the indirect selection phase, the corresponding superficially understood choice is selected together, maintaining coherence between them.

Nonexistent Entity The last choice arises when the model interprets the text correctly but fails to adequately consider the details of the image, resulting in a plausible yet incorrect choice. This resembles the correct answer in structure, but replaces the key object in the solution with one that does not exist in the image. To prevent the task from becoming too easy by generating highly irrelevant objects, we constrain the substituted object to one that, while absent from the image, is highly expected to be present in the scene and could replace the original entity. To identify such entities, the model is provided with the image as input to choose objects that align with the scene's context. This method ensures that the counterfactual choice leverages the expected coherence between the scene and its potential entities while rigorously testing the model's attention to visual details.

5. Experiments

Models We use the following models in our experiments. The detailed descriptions of each model are in Appendix C.

- Phi3.5-Vision-Instruct (4B) [29]
- LLaVA Onevision (7B) [23]
- Qwen2.5-VL-Instruct (7B, 72B) [43]
- InternVL-2.5-MPO (8B, 26B) [6]
- Idefics2 (8B) [22]
- LLaVA NeXT Vicuna (13B) [24]
- Ovis2 (16B) [28]
- GPT-4o [31]
- Gemini 1.5 Pro [12]
- InternVL-3 (38B) [42]

5.1. MLLMs Benefit from Visual Cues

Our first objective is to assess how effectively MLLMs leverage visual cues to resolve ambiguity in utterances. To this end, we systematically control the level of detail in the visual cues provided to the models and measure their accuracy in inferring the speaker's true intent. Performance is evaluated in both multiple-choice and free-form settings. For clarity, we primarily report multiple-choice accuracy, deferring free-form results to Appendix I.

We consider three levels of visual cues:

- *Language Models (LMs)* receive no visual input, requiring models to rely on superficial textual priors such as common-sense knowledge of sarcasm or humor to determine intent.
- *Socratic Models (SMs)* [47] use text-only LMs but incorporate short image captions (up to two or three sentences) as additional input. This short generic caption may lack sufficient detail, which may be insufficient for accurately inferring intent. Each SM model generated its own image captions and used them in subsequent processing.
- *Visual Language Models (VLMs)* receive the raw image input, enabling a more direct interpretation of visual cues.

Results The results in Tab. 1 indicate that MLLMs can leverage visual cues, albeit to a limited extent, since SMs and VLMs consistently outperform LMs across all evaluated models. Additionally, more detailed visual input generally improves performance, with VLMs surpassing SMs in most cases, except in the case of proprietary models. This exception is further analyzed in Sec. 5.2. Both the VCR and Ego4D subsets exhibit similar performance trends, demonstrating the generalizability of our findings across both staged and real-world scenarios. Finally, the consistently low performance of LMs further reinforces the validity of our dataset as a multimodal benchmark.

5.2. Analysis on Failure Modes

Here, we examine the ways in which models fail to infer the true intent and how these failure patterns vary with the level of visual cues provided. As shown in Fig. 3, our multiple-choice questions include three distinct types of incorrect answer candidates. We assess the model's *raw visual understanding* using the Fake Scene Understanding (FS) and Nonexistent Entity (NE) candidates, which test whether the model can correctly interpret the scene without being misled by fabricated or nonexistent elements. Conversely, the Superficial Understanding (SU) candidate evaluates the model's *reasoning ability*, testing whether it can go beyond surface-level perception to infer intent. We provide the full table of failure modes in the MCQ setting in Appendix I

Results Figure 4 illustrates how frequently each model selects different types of incorrect answers instead of the correct intent. Among the error types, Superficial Understanding (SU) is the most common. This indicates that while models generally succeed in recognizing basic visual details, they often fail to *reason* deeply about those visual cues to accurately infer the underlying intent of the speaker. However, proprietary models exhibit fewer SU-related errors, indicating stronger reasoning capabilities.

Moreover, stronger visual cues improve accuracy across all models and failure types. This improvement highlights the crucial role of visual conditioning in reducing both vision-based errors (FS and NE) and reasoning-related failures (SU).

Model	VAGUE-VCR			VAGUE-Ego4D		
	LM (L)	SM (L+V)	VLM (L+V)	LM (L)	SM (L+V)	VLM (L+V)
Phi3.5-Vision-Instruct (4B)	26.6	35.3 (↑ 8.7)	46.0 (↑ 19.4)	22.5	31.1 (↑ 8.6)	42.4 (↑ 19.9)
LLaVA-Onevision (7B)	13.1	29.4 (↑ 16.3)	43.1 (↑ 30.0)	11.3	29.5 (↑ 18.2)	43.2 (↑ 31.9)
Qwen2.5-VL-Instruct (7B)	11.1	25.6 (↑ 14.5)	46.8 (↑ 35.7)	9.8	28.0 (↑ 18.2)	48.4 (↑ 38.6)
InternVL-2.5-MPO (8B)	23.0	48.4 (↑ 25.4)	63.9 (↑ 40.9)	24.2	54.0 (↑ 29.8)	66.8 (↑ 42.6)
Idefics2 (8B)	13.9	21.1 (↑ 7.2)	58.7 (↑ 44.8)	14.8	18.2 (↑ 3.4)	58.3 (↑ 43.5)
LLaVA-NeXT-vicuna (13B)	24.2	37.2 (↑ 13)	46.4 (↑ 22.2)	20.3	34.1 (↑ 13.8)	52.5 (↑ 32.2)
Ovis2 (16B)	21.9	23.8 (↑ 1.9)	24.5 (↑ 3.6)	20.5	25.3 (↑ 4.8)	25.7 (↑ 5.2)
InternVL-2.5-MPO (26B)	21.2	48.5 (↑ 27.3)	63.7 (↑ 42.5)	21.8	55.2 (↑ 33.4)	68.7 (↑ 46.9)
InternVL-3 (38B)	24.8	47.2 (↑ 22.4)	63.6 (↑ 38.8)	18.0	47.5 (↑ 29.5)	59.8 (↑ 41.8)
Qwen2.5-VL-Instruct (72B)	29.6	55.6 (↑ 26.0)	**74.2 (↑ 44.6)**	26.8	59.3 (↑ 32.5)	**69.8 (↑ 43.0)**
GPT-4o	**46.4**	**69.5 (↑ 23.1)**	65.1 (↑ 18.7)	**48.2**	**67.5 (↑ 19.3)**	63.6 (↑ 15.3)
Gemini-1.5-Pro	43.2	62.4 (↑ 19.2)	60.6 (↑ 17.4)	40.3	60.6 (↑ 20.3)	60.6 (↑ 20.3)

Table 1. Experiments on the Multimodal Intention Disambiguation (MID) task with varying levels of visual cues. We report the accuracy (%) of the Multiple-Choice Question. ↑ indicates the performance gain from visual cues, i.e. increment compared to the LM setting.(L) denotes the use of language input only, while (L+V) indicates the incorporation of visual cues. The noticeable increase in accuracy across LM, SM, and VLM demonstrates that the introduction of detailed visual cues is beneficial for the task.

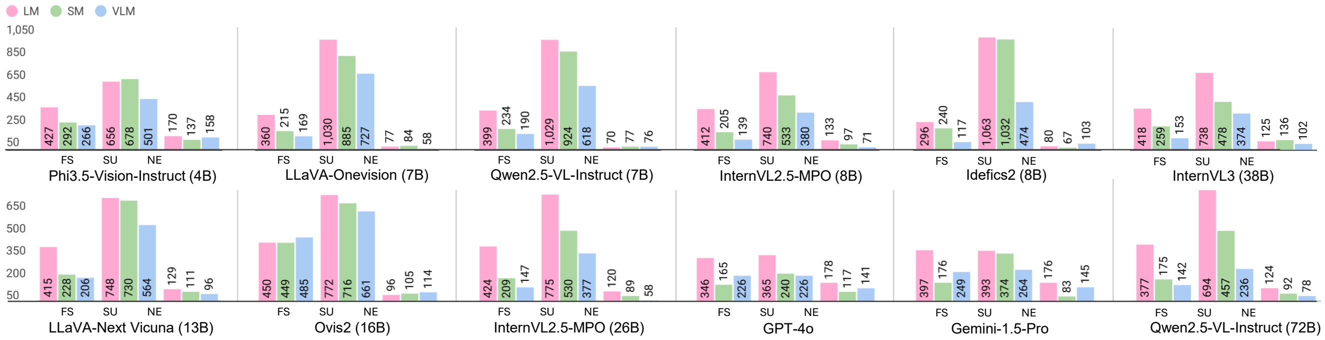


Figure 4. We present a bar plot to analyze the distribution of incorrect answer choices selected by each model. Each number represents how frequently a given choice was selected from the 1,677 items in the dataset. The counterfactual choice categories are FS (Fake Scene Understanding), SU (Superficial Understanding), and NE (Nonexistent Entity). We use distinct colors to represent LM (Language Models), SM (Socratic Models), and VLM (Visual-Language Models).

Notably, proprietary models perform better with captioned inputs (SM) than with raw images (VLM). A closer examination of Fig. 4 reveals that this discrepancy arises from vision-based failures (FS and NE) rather than reasoning-centric failure (SU). This indicates that their captioning ability potentially allows them to obtain more sophisticated information while reducing hallucination in the images. The supporting experiments and explanations are provided in Appendix F.

5.3. Comparison with human

To validate our benchmark and establish an upper bound for performance, we assess human accuracy, highlighting the gap between existing models and human capability. This evaluation follows the multiple-choice setup within the VLM setting, using a subset of 400 samples. As shown in Tab. 2, human performance reaches 94%, demonstrating near-perfect accuracy. Although proprietary models and some large-sized models show a decent performance, a notable performance gap (∼20%) still exists compared to human evaluators. This result underscores the significant gap between AI models' multimodal reasoning capabilities and human-level understanding when inferring hidden intent, suggesting that visual perception alone, even when accurate, is insufficient without deeper cognitive integration. This performance gap is due to the models' tendency to rely on surface-level text rather than understanding deeper visual-textual implications. Thus, advancing multimodal reasoning likely requires models to integrate higher-order cognitive processes, like commonsense reasoning and pragmatic understanding, into visual interpretation tasks. Refer

Model	Acc (%)	FS	SU	NE	Correct
Ovis2 (16B)	23.0	119	171	18	92
LLaVA-Onevision (7B)	41.0	43	183	10	164
Phi3.5-Vision-Instruct (4B)	44.3	60	132	31	177
Qwen2.5-VL-Instruct (7B)	47.0	47	152	13	188
LLaVA-NeXT-icuna (13B)	48.0	48	143	17	192
Idefics2 (8b)	57.0	28	120	24	228
Gemini-1.5-Pro	60.3	60	73	26	241
InternVL-3 (38B)	61.5	38	94	22	246
InternVL-2.5-MPO (8B)	61.8	42	95	16	247
GPT-4o	62.3	61	63	27	249
InternVL-2.5-MPO (26B)	63.0	36	101	11	252
Qwen2.5-VL-Instruct (72B)	72.3	39	60	12	289
Human	**94.0**	12	4	8	374

Table 2. Performance across models and humans on a sampled set of 400 questions. The results show that humans outperform models by a margin of over 20%.

Model	Type	Acc (%)	Incorrect count		
			FS	SU	NE
GPT-4o	SM	68.9	165	240	117
	SM+CoT	69.5 (↑ 0.6)	165	241	105
	VLM	64.6	226	226	141
	VLM+CoT	66.4 (↑ 1.8)	162	156	85
Gemini-1.5-Pro	SM	61.8	176	374	83
	SM+CoT	61.0 (↓ 0.8)	190	367	94
	VLM	60.6	249	264	145
	VLM+CoT	64.4 (↑ 3.8)	213	267	117

Table 3. Result of Chain-of-Thought (CoT) experiments on proprietary models, in both SM and VLM settings. ↑ and ↓ indicate an increase and decrease in accuracy when zero-shot CoT is applied.

to Appendix E for details on the selected subset and human evaluation setup.

5.4. Chain-of-Thought Experiments

Given the strong reasoning demands of multimodal intent deduction, we further explore the effectiveness of Chain-of-Thought (CoT) prompting [44] in enhancing the reasoning capabilities of MLLMs. The CoT prompt templates, provided in Fig. J21 and Fig. J22, are designed to explicitly ground the reasoning process, reducing hallucinations. While our main results focus on proprietary models, owing to their superior suitability for zero-shot CoT reasoning, we also present experiments and analyses for open-source models ranging from 4B to 72B in Appendix G.

Results As shown in Tab. 3, CoT prompting improves performance for raw image inputs (VLM), while showing no clear trend and remaining at a similar level for image caption inputs (SM). One possible explanation for this discrepancy is that CoT primarily enhances reasoning by improving grounding and reducing hallucinations. Since image captions inherently contain fewer hallucinations, SMs see some benefit from CoT prompting albeit at the cost of reduced detail. Additionally, the performance improvements observed with CoT prompting are consistent across different types of false answer candidates, suggesting a generalizable effect in enhancing reasoning quality.

6. Conclusion

We present VAGUE (Visual Contexts ClArify ambiGUous Expressions), a new benchmark aimed at assessing models' ability to interpret nuanced communication in complex multimodal scenarios. Our results show that models benefit from visual information when inferring the underlying intention of indirect expressions, as evidenced by their improved performance with increasing levels of visual cues. However, a significant disparity persists between machine capabilities and humans. To gain deeper insights into model inaccuracies, we design multiple-choice questions that explicitly address failure points, enabling a systematic and quantitative evaluation of the reasons behind performance. The primary challenge identified is the tendency of multimodal models to inadequately integrate visual cues, relying instead on the literal interpretation of textual information. This shortcoming highlights the need for further research, and we anticipate that VAGUE will open up promising avenues for developing systems capable of deeper multimodal reasoning to enhance AI's ability to engage in human-like interactions.

7. Limitations

First, we acknowledge that there are cultural and linguistic limitations. We incorporate sarcastic, humorous, and idiomatic implicatures in generating indirect expressions. However, since the initial data drafts are created by a model (GPT-4o [31]), they may reflect cultural biases present in its training data. To prevent any potential ethical issues, all human annotators are instructed to remove any content considered problematic or discriminatory during the rating and filtering process. Also, all textual expressions in VAGUE are limited to English. Therefore, we encourage future exploration of indirect expressions across diverse languages and cultures. The second limitation pertains to the dependency of certain meta-information on the quality of the parent dataset and the performance of the models utilized during the dataset generation. We observe that bounding box annotations from YOLOv11 [18] are occasionally duplicated, leading to a reported number of people higher than actually present. Likewise, the tagging model RAM [49] sometimes misidentified objects. Therefore, we remove any corrupted instances that could undermine the integrity of the task during the human rating and filtering process.

References

[1] Ashwin Baluja. Text is not all you need: Multimodal prompting helps llms understand humor. *arXiv preprint arXiv:2412.05315*, 2024. 3

[2] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, and Aditya Ramesh. Improving image generation with better captions. 1

[3] Matteo Bortoletto, Constantin Ruhdorfer, Lei Shi, and Andreas Bulling. Explicit modelling of theory of mind for belief prediction in nonverbal social interactions. In *Proc. 27th European Conference on Artificial Intelligence (ECAI)*, pages 1–8, 2024. 2

[4] Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. 2

[5] Zhawnen Chen, Tianchun Wang, Yizhou Wang, Michal Kosinski, Xiang Zhang, Yun Fu, and Sheng Li. Through the theory of mind's eye: Reading minds with multimodal video large language models. *arXiv preprint arXiv:2406.13763*, 2024. 2

[6] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 6, 3

[7] Christine Cuskley, Rebecca Woods, and Molly Flaherty. The limitations of large language models for understanding human language and cognition. *Open Mind*, 8:1058–1083, 2024. 2

[8] Hang Du, Guoshun Nan, Sicheng Zhang, Binzhu Xie, Junrui Xu, Hehe Fan, Qimei Cui, Xiaofeng Tao, and Xudong Jiang. Docmsu: A comprehensive benchmark for document-level multimodal sarcasm understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 17933–17941, 2024. 2

[9] Jiafei Duan, Samson Yu, Nicholas Tan, Li Yi, and Cheston Tan. Boss: A benchmark for human belief prediction in object-context scenarios. *arXiv preprint arXiv:2206.10665*, 2022. 2

[10] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. *CVPR 2024*, 2024. 4

[11] Kanishk Gandhi, Jan-Philipp Fraenkel, Tobias Gerstenberg, and Noah D. Goodman. Understanding social reasoning in language models with language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*, 2023. Dataset and Benchmarks Track. 2

[12] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 6, 3

[13] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. UR-FUNNY: A multimodal language dataset for understanding humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2046–2056, Hong Kong, China, 2019. Association for Computational Linguistics. 2

[14] Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[15] EunJeong Hwang and Vered Shwartz. Memecap: A dataset for captioning and interpreting memes. *arXiv preprint arXiv:2305.13703*, 2023. 3

[16] Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. *arXiv preprint arXiv:2401.09899*, 2024. 3

[17] Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[18] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. 4, 8

[19] Shivani Kumar, Atharva Kulkarni, Md Shad Akhtar, and Tanmoy Chakraborty. When did you become so smart, oh wise one?! sarcasm explanation in multi-modal multi-party dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5956–5968, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[20] Shivani Kumar, Ishani Mondai, Md Shad Akhtar, and Tanmoy Chakraborty. Explaining (sarcastic) utterances to enhance affect understanding in multimodal dialogues. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press, 2023. 2

[21] Yash Kumar Lal and Mohaddeseh Bastan. SBU figures it out: Models explain figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 143–149, Abu Dhabi, United Arab Emirates (Hybrid), 2022. Association for Computational Linguistics. 2

[22] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language mod-

els? *Advances in Neural Information Processing Systems*, 37: 87874–87907, 2025. 6, 3

[23] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 3

[24] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 6, 3

[25] Kevin Y. Li, Sachin Goyal, Joao D. Semedo, and J. Zico Kolter. Inference optimal vlms need fewer visual tokens and more parameters, 2025. 4

[26] 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*, pages 740–755, Cham, 2014. Springer International Publishing. 1, 3

[27] Annie Louis, Dan Roth, and Filip Radlinski. "I'd rather just go to bed": Understanding indirect answers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7411–7425, Online, 2020. Association for Computational Linguistics. 2

[28] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024. 6, 3

[29] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone, 2024. 6, 3

[30] Maryam Sadat Mirzaei, Kourosh Meshgi, and Satoshi Sekine. What is the real intention behind this question? dataset collection and intention classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13606–13622, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[31] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 5, 6, 8, 3

[32] Kun Ouyang, Liqiang Jing, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. Sentiment-enhanced graph-based sarcasm explanation in dialogue. *arXiv:2402.11414*, 2024. 2

[33] 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*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 3, 4

[34] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4): 515–526, 1978. 2

[35] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017. 4

[36] Melanie Sclar, Sachin Kumar, Peter West, Alane Suhr, Yejin Choi, and Yulia Tsvetkov. Minding language models' (lack of) theory of mind: A plug-and-play multi-character belief tracker. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 13960–13980, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[37] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multi-modal multi-agent theory of mind, 2024. 2

[38] 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*, pages 12075–12097, Bangkok, Thailand, 2024. Association for Computational Linguistics. 2

[39] Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. DeepMet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 30–39, Online, 2020. Association for Computational Linguistics. 2

[40] Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. DIRECT: Direct and indirect responses in conversational text corpus. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1980–1989, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2

[41] Ego4d Team. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012, 2022. 2, 4

[42] OpenGVLab Team. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. 6, 3

[43] Qwen Team. Qwen2.5-vl technical report, 2025. 6, 3

[44] 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. 8

[45] Alex Wilf, Leena Mathur, Sheryl Mathew, Claire Ko, Youssouf Kebe, Paul Pu Liang, and Louis-Philippe Morency. Social-iq 2.0 challenge: Benchmarking multimodal social understanding. `https://github.com/abwilf/Social-IQ-2.0-Challenge`, 2023. 2

[46] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6713–6724, 2018. 2, 4, 1

[47] Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 6

[48] 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. 4

[49] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, Yandong Guo, and Lei Zhang. Recognize anything: A strong image tagging model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1724–1732, 2024. 4, 8, 1