VAGUE: Visual Contexts Clarify Ambiguous Expressions

Supplementary Material

A. Directness & Indirectness Examples

Figure J3 provides a more concrete explanation with realistic examples of the direct and indirect expressions defined in the main text. Each expression follows two key criteria, and we avoid examples like those in the "Bad" column while prioritizing those in the "Good" cases.

B. VAGUE Benchmark

This section introduces details of VAGUE Benchmark dataset. We provide examples of images and their corresponding multiple-choice questions, along with the prompts used to generate direct, indirect expression, correct understanding expression and superficial understanding expression. Additionally, we present the prompts used to generate two incorrect answer choices of multiple choices set (fake scene understanding expression, nonexistent entity expression) and the human rating criteria employed to assess the quality of direct and indirect expressions.

B.1. Samples in VAGUE

Figures J5 to J7 illustrate six examples from our benchmark dataset. In the Visual Language Models (VLMs) setting, the model is presented with an image containing person indicator tags, a question, and the speaker's indirect utterance, as shown in the figures, to answer a multiple-choice question. For reference, we have included the corresponding direct expression below each sample.

B.2. Benchmark Statistics

Table B1 illustrates the average statistics of the datasets that make up VAGUE-VCR and VAGUE-Ego4D. "Average object counts" refers to the average number of detected objects per image, while "Average people counts" indicates the average number of detected individuals per image. "Average word counts" represents the average number of words in direct and indirect expressions generated for each image. Notably, the high values of "Average object counts" and "Average people counts" suggest that the images are not simplistic.

	VAGUE-VCR	VAGUE-Ego4D
Average object counts	7.4	6.89
Average people counts	4.48	2.59
Average word counts (direct)	9.69	15.9
Average word counts (indirect)	11.52	12.27

Table B1. Dataset statistics table

Furthermore, Fig. B1 illustrates the diversity of intentions generated from our two parent datasets. Using the 20 most frequently occurring verbs in the solution triplets (person, action, object) of each dataset, we generate a radial diagram. Both VAGUE-VCR and VAGUE-Ego4D exhibited a comparable level of diversity, demonstrating that while not perfectly uniform, the dataset covers a wide range of contexts.

B.3. Details of Object Extraction

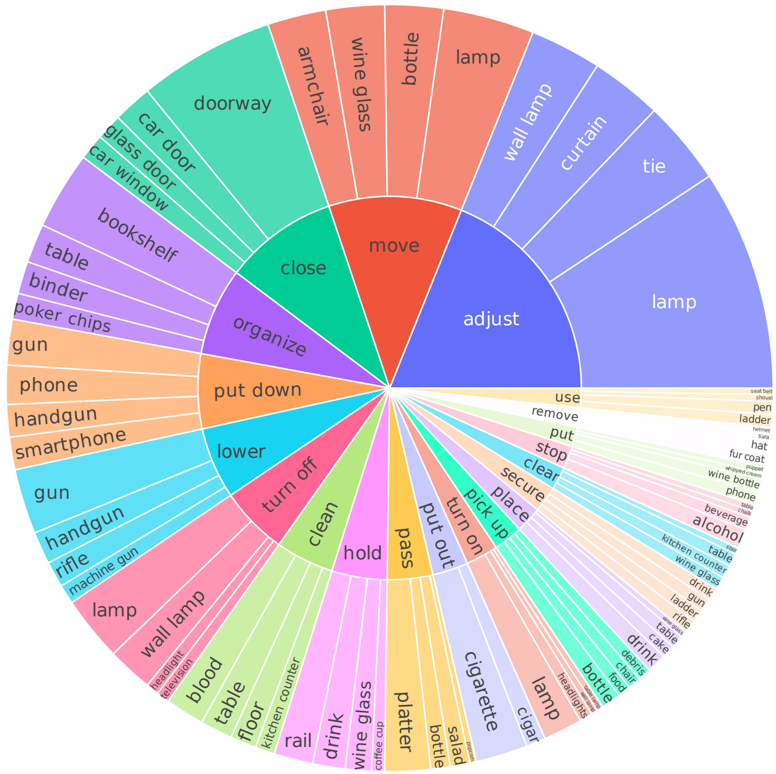
Our prompts, both direct and indirect, are crafted as instructions that request the recipient to perform specific manipulations on an object within the scene. Such formulation of the task prompt requires the scene to have enough objects. Although the VCR [46] dataset contains COCO [26] object tags as meta-information, COCO objects are highly limited and often fail to comprehensively capture the objects present in real-world scenes. Therefore, we process our images using the Recognize Anything Model (RAM) [49] to identify physical objects in each image. However, RAM [49] frequently generates tags for entities that are not strictly physical objects, such as places, emotions, and colors. To address this, we manually curate a list of 2,403 physical objects from the full set of 4,585 items detectable by RAM [49]. Using this refined list, we filter the initially extracted entities from the images for further usage.

B.4. OCR Experiments for Testing Person Tags

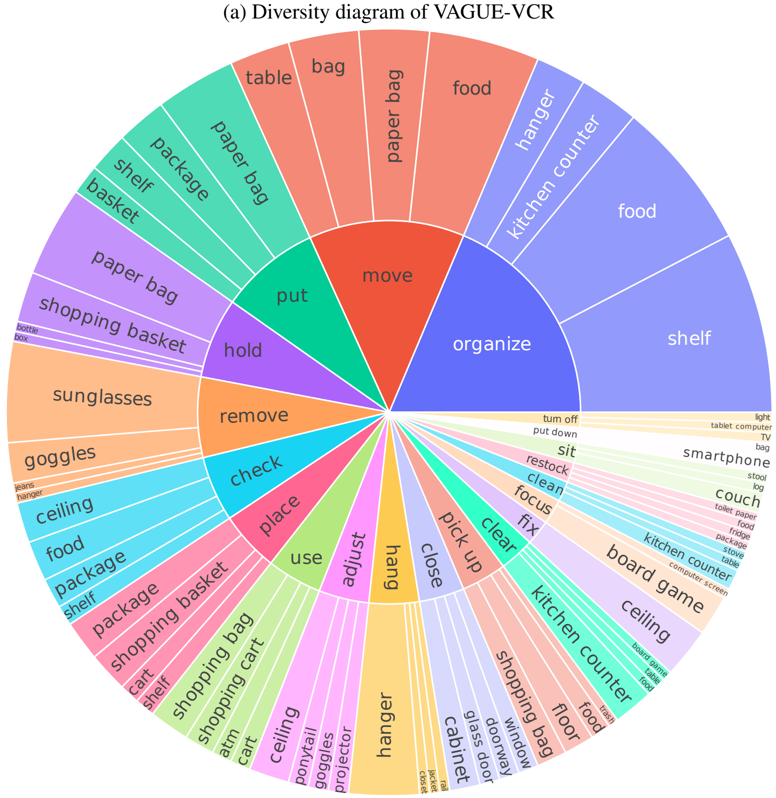
In Sec. 4.1.3, we incorporate person indicators into images to distinguish individuals when interpreting prompts in various contexts. While this provides a straightforward method for grounding the target person, it requires models to perform basic Optical Character Recognition (OCR) to interpret phrases such as "Hey person2" To evaluate this capability, we select three images from the COCO [26] dataset, each containing a different number of people. Fig. B2 presents these images, featuring two, five, and ten individuals wearing t-shirts in various colors. By asking the models to identify the t-shirt color of specific individuals, we conclude that the selected models consistently performed perfectly in recognizing person indicators as shown in Tab. B2

B.5. Model Instruction for Direct and Indirect prompts

Fig. J8 is the prompt used to generate direct expressions. Also, Fig. J9 is a prompt that generates the correct answer for multiple choice by understanding the intention based on this direct expression(Correct). Fig. J10 is the prompt



(a) Diversity diagram of VAGUE-VCR



(b) Diversity diagram of VAGUE-Ego4D

Figure B1. Diversity diagrams of the 20 most frequent actions in VAGUE-VCR dataset and VAGUE-Ego4D dataset respectively.

Table B2

model	response		
	(a)	(b)	(c)
Phi3.5-Vision-Instruct (4B)	Person1 is wearing a red shirt.	Person1 is wearing a red shirt.	Person5 is wearing a red shirt.
LLaVA-Onevision (7B)	red	red	red
Qwen2.5-VL-Instruct (7B)	Person1 is wearing a red shirt.	Person1 is wearing a red shirt.	Person5 is wearing a red shirt.
InternVL-2.5-MPO (8B)	Person 1 is wearing a red shirt.	Person 1 is wearing a red shirt.	Person 5 is wearing a red shirt.
Idefics2 (8B)	Red.	Red.	Red.
LLaVA-NeXT-vicuna (13B)	Person 1 is wearing a red t-shirt.	Person 1's t-shirt is red.	Person 5 is wearing a red t-shirt.
Ovis2 (16B)	Person1 is wearing a red shirt.	Person1 is wearing a red shirt.	Person5 is wearing a red shirt.
InternVL-2.5-MPO (26B)	Person 1 is wearing a red shirt.	Person 1 is wearing a red shirt.	Person 5 is wearing a red shirt.
InternVL-3 (38B)	Person 1 is wearing a red shirt.	Person 1 is wearing a red shirt.	Person 5 is wearing a red shirt.
Qwen2.5-VL-Instruct (72B)	Person 1 is wearing a red shirt.	Person 1 is wearing a red shirt.	Person5 is wearing a red shirt.
GPT-4o	Person 1 is wearing a red shirt.	Person 1 is wearing a red shirt.	Person 5 is wearing a red shirt.
Gemini-1.5-Pro	Person 1 is wearing a red t-shirt.	Person 1 is wearing a red t-shirt.	Person 5 is wearing a red t-shirt.

(a) (b) (c)

Figure B2. Three images from the COCO [26] dataset which were used assessing OCR capability. We asked about the t-shirt color of person 1 in (a) and (b), and person 5 in (c). The correct answer, "red," was identified correctly by all selected models during testing.

that generates indirect expressions based on direct expressions and their intended meaning. Additionally, it produces superficially misinterpreted intentions (SU).

B.6. Human Rating for Direct and Indirect prompts

This section details the human rating and filtering process. Human annotators are instructed to follow the scoring guidelines provided in the tables in Figs. J13 and J14. To facilitate high-quality annotations efficiently, we use *Label Studio* (`https://labelstud.io/`). Figure J13 illustrates the UI for rating direct expressions, while Fig. J14 shows the UI for selecting and rating indirect expressions.

B.7. Model Instruction for Counterfactual Choices

Fig. J11 is the prompt used to generate incorrect answer choices by creating fake captions that aligns with the indirect expressions but is inconsistent with the direct expressions and combining fake captions with the indirect expressions to derive the plausible intention (FS). Fig. J12 is the prompt that generates incorrect answer choices by introducing objects not present in the image, intentionally misrepresenting the intended meaning of the indirect expression (NE).

C. Baseline Models

To ensure broad coverage of existing multimodal language models (MLLMs), we evaluate ten open-source models with varying parameter sizes alongside two closed-source models. For the open-source models, we use Phi3.5-Vision-Instruct (4B) [29], optimized for concise instruction-following tasks, providing robust text-image alignment in low-parameter settings. LLaVA Onevision (7B) [23] focuses on high-resolution image interpretation, enhancing multimodal dialogue through refined attention mechanisms. Qwen2.5-VL-Instruct (7B, 72B) [43] uses advanced vision-language pretraining to handle diverse image-based queries and textual instructions. InternVL-2.5-MPO (8B, 26B) [6] is designed for multi-purpose optimizations, supporting enhanced multimodal reasoning. InternVL-3 (38B) [42] provides better performance in various tasks than in the previously released models. Idefics2 (8B) [22] adopts a compact architecture for efficient training, emphasizing domain-specific image understanding and textual generation. LLaVA NeXT Vicuna (13B) [24] employs an improved vision encoder and refined instruction tuning for enhanced commonsense reasoning. Ovis2 (16B) [28] excels in image captioning and inference, driven by robust textual grounding and visual alignment. Among proprietary models, GPT-4o [31] demonstrates advanced language comprehension paired with visual perception for nuanced multimodal interactions. Gemini 1.5 Pro [12] integrates high-resolution vision processing with a powerful language model, delivering refined instruction-following and cross-domain reasoning.

D. Free-From Answering

D.1. Metrics

BLEU [33] The BLEU score is a metric for assessing the quality of machine-generated text by comparing it to a reference. It measures n-gram precision, checking how many n-grams from the generated text appear in the reference. However, when evaluating free-form generated text with BLEU, the score drops from 2-grams onward due to the high variability in phrasing. To obtain a more meaningful measure, we use 1-gram BLEU, which captures individual word overlap and provides a reasonable approximation of text similarity.

BERT-F1 [48] BERT-F1 is a semantic similarity metric that utilizes contextual embeddings from the BERT model. Instead of relying on exact word matches, it calculates an F1-score based on token similarity in an embedding space. This allows it to capture paraphrasing and synonymy, making it more effective at evaluating meaning rather than just surface-level similarity.

D.2. Limitations of Free-form Answering

Traditional text similarity metrics such as BLEU [33] and BERT-F1 [48] are widely used for evaluating language models. However, they are often not well-suited for assessing intent similarity in multimodal intent disambiguation tasks.

BLEU computes n-gram overlap between sentences, treating all tokens equally, which makes it ineffective in capturing subtle intent differences. Similarly, BERT-F1, which uses contextual embeddings, struggles with antonymy, as opposing words often appear in similar contexts. For example, *"open the window"* and *"close the window"* have completely opposite intents, yet BERT-F1 assigns them a high similarity score of 0.939 due to shared structure and overlapping words.

To address this, we adopt a multiple-choice question (MCQ) format as our main setting, which directly evaluates whether a model selects the correct intent rather than relying on approximate similarity scores. The MCQ structure explicitly includes plausible but incorrect distractors, ensuring that models must resolve ambiguity by integrating multimodal cues rather than relying on lexical overlap alone. This structured approach enables a more robust and intent-aware evaluation.

D.3. Qualitative Example

Fig. J4 shows the results of free-form answering in the VLM, SM, and LM settings using the InternVL-2.5-8B-MPO model. It shows that the underlying intention behind the indirect expression in the given image is well preserved, and the generated responses across different settings are highly similar.

E. Details of Human Evaluation

For human evaluation, we use a subset of 400 high-quality items from VAGUE. These items are carefully selected to ensure that the intended correct answer aligns well with the image and that the indirect expression remains sufficiently ambiguous. Low-quality samples have already been filtered out, resulting in direct expressions with scores of 4 or 5 and indirect expressions with scores ranging from 3 to 5. Due to this filtering, the average scores of direct and indirect expressions are often tied. In such cases, we randomly select items to form the final 400-item subset.

Our human evaluator is a student researcher with expertise in human cognition across modalities and language models. The evaluator, a Korean fluent in English at a native level, annotated all 400 items independently.

F. Why does SM setting outperform in proprietary models?

Table F3. ↑ denotes MCQ performance boost in SM when switching to GPT-4o-generated captions.

Models	Self captioning	GPT-4o captioning
Phi3.5-Vision-Instruct (4b)	34.0	37.9 (↑ 3.9)
LLaVA-Onevision (7b)	29.4	36.5 (↑ 7.1)
LLaVA-NeXT-vicuna (13B)	36.3	41.7 (↑ 5.4)
InternVL-2.5-MPO (26B)	50.6	50.8 (↑ 0.2)

The rationale benind the superior SM performance of proprietary models lies in their reduced vision-related (FS, NE) errors. We elaborate this in two points: 1) If captions lack necessary detail, VLMs naturally excel. Tab. F3 implies proprietary models have better captions, by showing a performance boost when leveraging GPT-4o's captions in smaller open-source models within the SM setting. 2) If the caption somehow attains *parity of detail* (Practically, length constraints preclude full nuance capture.) with the image, text tokens handle visual information more effectively than image tokens [25]. Indeed, we observed in CoT reasoning that VLMs lean more on the speaker's rationale while SMs take a more scene-focused approach

G. CoT Open-source

We provide CoT ablations on open-source models and compare them to our SM and VLM baselines. Tab. G4 shows that CoT slightly degrades SM in the most cases, while VLMs experience a much bigger drop in their performance, increasing all types of errors (FS, SU, NE) together. Consistent degradation shows open-source models lag proprietary ones in deep reasoning. We believe the sharper drop in VLMs stems from *conditioning dilution* [10] due to longer generation sequences.

H. Model Instruction for Experiments

This section presents the prompts used for experiments involving different settings for each model: *Visual Language*

Model	Type	Acc (%)	Incorrect count		
			FS	SU	NE
Phi3.5-Vision-Instruct (4B)	SM	34.0	292	678	137
	SM+CoT	31.5 (↓ 2.5)	291	603	148
	VLM	44.8	266	501	158
	VLM+CoT	36.6 (↓ 8.2)	295	510	154
LLaVA-Onevision (7B)	SM	29.4	215	885	84
	SM+CoT	28.0 (↓ 1.4)	291	795	101
	VLM	43.1	169	727	58
	VLM+CoT	33.1 (↓ 10.0)	187	736	76
LLaVA-NeXT-vicuna (13B)	SM	36.3	228	730	111
	SM+CoT	32.6 (↓ 3.7)	259	598	149
	VLM	48.4	206	564	96
	VLM+CoT	38.3 (↓ 10.1)	206	541	139
InternVL-2.5-MPO (26B)	SM	50.6	209	530	89
	SM+CoT	45.1 (↓ 5.5)	213	584	106
	VLM	65.3	147	377	58
	VLM+CoT	58.1 (↓ 7.2)	17	440	67
InternVL-3 (38B)	SM	47.3	259	478	136
	SM+CoT	54.3 (↑ 7.0)	166	359	102
	VLM	62.4	153	374	102
	VLM+CoT	52.2 (↓ 10.2)	211	387	107
Qwen2.5-VL-Instruct (72B)	SM	56.8	175	457	92
	SM+CoT	55.8 (↓ 1.0)	213	428	100
	VLM	72.8	142	236	78
	VLM+CoT	69.9 (↓ 2.9)	154	272	78

Table G4. Result of Chain-of-Thought (CoT) experiments on open-source models, in both SM and VLM settings. ↑ and ↓ indicate an increase and decrease in accuracy when zero-shot CoT is applied.

Models (VLMs), *Socratic Models (SMs)*, and *Language Models (LMs)* in both multiple-choice questions and free-form answering tasks. Additionally, it provides the full results obtained from these experiments.

Multiple Choice Questions Fig. J15, Fig. J16, Fig. J17 show the prompts used for multiple-choice question experiments under VLM, SM, and LM settings.

Free-form Answering Fig. J18, Fig. J19, Fig. J20 show the prompts used for free-form answering experiments under VLM, SM, and LM settings.

Chain-of-Thoughts Additionally, Fig. J24, Fig. J22 are prompts that we use for zeroshot chain-of-thought experiments with Socratics Models, while Fig. J23, Fig. J21 are those with Visual-Language Models.

I. Full Results

Table J5 and Table J6 present the complete experimental results conducted in this paper for the VAGUE-VCR and VAGUE-Ego4D datasets, respectively. The total number of items in VAGUE-VCR is 1,144, and in VAGUE-Ego4D, it is 533. However, the item count in the *valid count* column does not always match these totals. This discrepancy occurs when the model responds with 'I don't know' or refuses to answer. When calculating accuracy, we treat such cases as incorrect and divide the number of correct responses by the total number of items in each dataset.

J. Full structure of VAGUE

Fig. J25 shows the structure of our benchmark dataset, VAGUE.

		Bad	Good
Direct	Relevance	**Irrelevant** Unrelated to visual. *It is* ***impossible*** *to draw a connection between the text prompt with the image.* e.g. (Image: A room without a window) Hey person1, please close that window.	**Relevant** Clearly related to visual. *There* ***clearly possible*** *to draw a connection between the prompt and image at first glance.* e.g. (Image: A room with an open window, with snow piled up outside.) Hey person1, please close that window because it's cold.
	Solvability	**Unsolvable** Mulitple independent solutions possible. *We can draw multiple* ***independent*** *solutions from the prompt.* e.g. Hey person2, 1) put the gun down immediately and 2) get behind the car for cover.	**Solvable** Single independent solution only. *We can draw a* ***single*** *clearly solvable incident.* e.g. Hey person3, please bring the refreshments over here faster.
Indirect	Consistency	**Inconsistent** Indirect intention ≠Direct intention ***No possible interpretation*** *of the indirect prompt matches that of the direct prompt.* e.g.(Direct : Hey person13, stop standing by the wall.) Hey person13, are we building an igloo here?	**Consistent** Indirect intention = Direct intention ***The intention*** *of the indirect prompt matches the intention of the direct prompt.* e.g. (Direct : Hey person3, listen to what person2 has to say!) Hey person3, I believe person2 has some important words to share with you.
	Ambiguity	**Direct** Replicated. *The prompt merely* ***rephrases*** *the direct prompt via* ***direct rephrasing.*** e.g. (Direct: Hey person2, stop distracting the horse while person 18 is talking.) Hey person2, maybe now isn't the best time to bond with our four-legged friend.	**Ambiguous** Well protected. *The prompt* ***almost perfectly ambiguates*** *the action to be performed and the entity to be adjusted.* e.g. (Direct: Hey person3, can you hurry up with the refrigerator?) Hey person3, is it just me or is the energy bill increasing lately?

Figure J3. This table presents two evaluation criteria for direct expressions and two for indirect expressions, along with descriptions and corresponding bad and good cases. The examples in the bad and good cases are derived from human ratings based on the given criteria.

[Free-form Answer] 0013_Halloween_00.15.15.492-00.15.17.652@0_annot.jpg

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

[Answers]

LM: The speaker wants person1 to notice the unusual nature of the parking situation.

SM: The speaker wants person1 to notice the special nature of the parking spot.

VLM: The speaker wants person1 to notice the special nature of the handicapped parking spot.

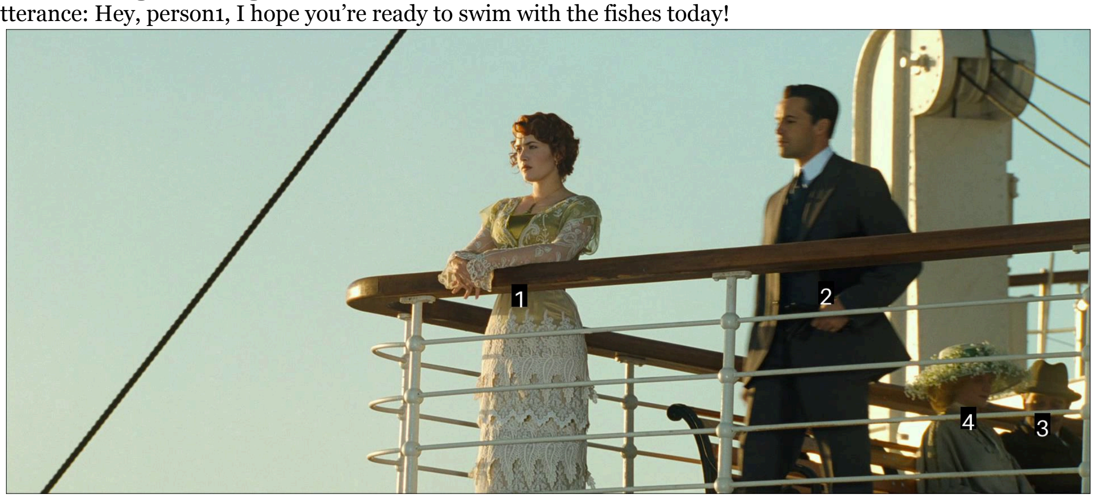
Figure J4. This figure shows the results generated by InternVL-2.5-8B-MPO for indirect expressions and free-form responses across VLM, SM, and LM settings.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey, person1, spot the difference, this parking's a bit too special, isn't it?

A) The speaker wants person1 to admire the usually decorated motorcycle in the parking lot. *(Fake Scene Understanding)*
B) The speaker wants person1 to enjoy playing a puzzle game and spot differences. *(Superficial Understanding)*
C) The speaker wants person1 to move the sedan because it's in a handicapped parking spot. *(Correct)*
D) The speaker wants person1 to move the sedan because it's parked in front of a fire hydrant. *(Nonexistent Entity)*

Direct expression (reference): Hey person1, you should move the sedan from the handicapped parking spot.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey, person1, I hope you're ready to swim with the fishes today!



A) The speaker wants person1 to use the safety harness to ensure they don't fall overboard. *(Nonexistent Entity)*
B) The speaker wants person1 to adjust their position and avoid stressing the rail. *(Correct)*
C) The speaker wants person1 to grasp the snorkel and prepare for diving. *(Fake Scene Understanding)*
D) The speaker wants Person1 to jump into the water and swim. *(Superficial Understanding)*

Direct expression (reference): Hey person1, please be careful not to lean too heavily on the rail.

Figure J5. These are examples of an image along with its corresponding generated direct expression and multiple-choice question set.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey, person1, did someone forget to pay the electricity bill?

A) The speaker wants Person1 to check if there was an oversight with paying the electricity bill. *(Superficial Understanding)*
B) The speaker wants person1 to check the candle for effective illumination in the room. *(Fake Scene Understanding)*
C) The speaker wants person1 to open the curtains to let in more natural light. *(Nonexistent Entity)*
D) The speaker wants person1 to adjust the lamp to enhance lighting on the dining table. *(Correct)*

Direct expression (reference): Hey person1, please adjust the lamp so it provides better lighting on the dining table.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey person2, feeling like a dragon today, aren't we?

A) The speaker wants person2 to adjust the flowing red scarf like a dragon's wings. *(Fake Scene Understanding)*
B) The speaker wants person2 to extinguish the cigarette to improve the atmosphere. *(Correct)*
C) The speaker wants Person2 to embrace their inner dragon and keep 'breathing fire.' *(Superficial Understanding)*
D) The speaker wants person2 to turn off the scented candle to improve the atmosphere. *(Nonexistent Entity)*

Direct expression (reference): Hey, person2, please put out the cigarette.

Figure J6. These are examples of an image along with its corresponding generated direct expression and multiple-choice question set.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey, person1, looks like you're planning a party for the neighborhood squirrels.

A) The speaker wants person1 to describe the tiny picnic tables' setup for squirrels. *(Fake Scene Understanding)*
B) The speaker wants person1 to close the cereal box to keep it fresh. *(Correct)*
C) The speaker wants person1 to use a chip clip to seal the cereal box. *(Nonexistent Entity)*
D) The speaker wants person1 to entertain the neighborhood squirrels with a cereal party. *(Superficial Understanding)*

Direct expression (reference): Hey, person1, please close the cereal box to keep the contents fresh.

[Question] Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Utterance: Hey, person1, please close the fridge door to keep the food fresh.

A) The speaker wants person1 to build an igloo in the kitchen. *(Superficial Understanding)*
B) The speaker wants person1 to arrange the pillows neatly into a playful fortress. *(Fake Scene Understanding)*
C) The speaker wants person1 to use the oven mitts to handle hot items safely.*(Nonexistent Entity)*
D) The speaker wants person1 to close the fridge to preserve the food's freshness. *(Correct)*

Direct expression (reference): Hey, person1, please close the fridge door to keep the food fresh.

Figure J7. These are examples of an image along with its corresponding generated direct expression and multiple-choice question set.

Prompt for Direct Generation

Your job is to do two things.

1. Generate a direct complaint based on the image. Your generated prompt must keep these two criteria in mind:
    a. Specify the recipient: The speaker is the person who is viewing the scene. Specify the recipient as a person in the image (begins with "Hey, person1…"). There is a number tag in the image for each person.
    b. Generate direct prompts: Your complaint must include the "subject", "action", "object", and "reason": it should convey the "WHO should do WHAT action on WHAT object WHY."
2. Generate a solution triplet that addresses the prompt. Your generated solution must keep these three criteria in mind:
    a. Triplet: The format of your output must be in (subject, action, object).
    b. Problem Mitigation: The generated solution must address the prompt in a way that resolves the complaint in the prompt.
    c. Solvable with Physical object: The object from the triplet–(subject, action, object)–must be a physical object from the provided "Entity" list.

Entity: {entities}
Prompt: (One Statement)
Solution: (Subject, Action, Object)
Caption: (2-3 Sentences Describing the Scene)

Figure J8. This prompt selects one of the list of entities and generates a direct request to a person in the image. It also generates a triple solution and generates a caption for the scene.

Prompt for mcq-correct Generation

Your job is to figure out the speaker's true intention based on the given prompt. Your generated response must keep these three criteria in mind:
1. Your answer should include {action}(the action to execute) and {obj}(object of being manipulated).
2. You should answer to this specific prompt: {direct}
3. Your answer should not exceed 15 words and start sentence with 'The speaker wants'

Figure J9. This prompt takes an action, object, and direct expression as input and outputs the true intention behind the direct expression.

Prompt for Indirect, mcq-Superficial Understanding(SU) Generation

Your job is to rewrite the following sentence into three different indirect sentences and to think about the possible misinterpreted intention of each.
The likely misinterpreted intention of your generated sentence should not be clearly different from the original intention.
You will be given the original sentence, and original intention, as well as a scene description.

These are the requirements for the indirect sentence:
1. Indirectness: The prompt should be indirect, maybe slightly sarcastic, humorous, or even use an idiom to hide the true intention, as opposed to the direct version.
2. Object Absence: The prompt must not contain the “OBJECT” or “ACTION”, or anything similar or synonymous in any way, from the original intention.
3. Natural Communication: The prompt should be a simple and natural day-to-day conversational statement, not pedantic.

The generated indirect sentence should have a clearly different superficial meaning.
Very importantly, the likely misinterpreted intention should sound off in the given situation (when understood literally).

For example, “Hey Person1, please clean your room!” can become an indirect sentence:
“Hey Person1, this is a disaster!”
Here, the likely misinterpreted intention might be: “The speaker wants Person1 to escape from the disaster.”
This is clearly hilarious in the context of facing a messy room.

Like the example above, note that the likely misinterpreted intention should start with:
“The speaker wants Person N to ...”

Original sentence: {direct}
Original Intention: {correct}
PROHIBITED words in indirect sentences: {action}, {obj}, and synonyms of {obj}.
Scene description: {caption}

1. Indirect sentence (use sarcasm): Hey person{p}, (Your answer)
Likely misinterpreted intention (superficial understanding): The speaker wants person{p} to (Your answer)

2. Indirect sentence (user humor): Hey person{p}, (Your answer)
Likely misinterpreted intention (superficial understanding): The speaker wants person{p} to (Your answer)

3. Indirect sentence(use meme or slang): Hey person{p}, (Your answer)
Likely misinterpreted intention (superficial understanding): The speaker wants person{p} to (Your answer)

Figure J10. This prompt takes a direct expression, true intention, and caption as input to generate an indirect expression that conveys the intended meaning. Additionally, it produces a superficially interpreted version of the indirect expression.

Prompt for mcq-Fake Scene Understanding(FS) Generation

Your job is to guess the underlying intention of the speaker when the situation is given.
Your generated response must keep these 5 criteria in mind:
1. Current situation: {fake_caption}
2. You should answer to this specific prompt: {indirect}
3. Your answer should not exceed 15 words and start sentence with 'The speaker wants {p} to'
4. Your answer should talk about one physical object in that given situation.
5. Your answer SHOULD NOT HAVE THE SAME MEANING to "{answer}"

Example
a. {icls[0]}
b. {icls[1]}
c. {icls[2]}

Figure J11. This prompt uses a fake caption and an indirect expression to derive the most plausible intention that does not align with the true intention of direct expression.

Prompt for mcq-Nonexistent Entity(NE) Generation

You are given an indirect complaint of the situation portrayed in the image and a correct interpretation of the prompt.
Your job is to intentionally come up with an incorrect interpretations that will serve as an incorrect choice for a multiple choice question.
The incorrect interpretation should specifically be designed using an object that is reasonable but does not exist anywhere in the image.
The object, although not in the image, should be one that is very highly be expected to be present in the scene and could be used to resolve the complaint.

Your answer should not exceed 15 words and start sentence with 'The speaker wants'

[Example 1]
Prompt: Hey person1, I guess we're all going to be sharing more than just food today.
Interpretation: The speaker wants person1 to use a fork to pick the food from the platter.
Assistant: The speaker wants person1 to make use of the serving spoon when picking up shared food.

[Example 2]
Prompt: Hey person1, I guess we're all going to be sharing more than just food today.
Interpretation: The speaker wants person1 to use a fork to pick the food from the platter.
Assistant: The speaker wants person1 to use the plastic gloves to avoid sharing germs.

[You]
Prompt: {text}
Interpretation:{interpretation}

Figure J12. This prompt takes an indirect expression and an interpretation of the indirect expression as input to generate an incorrect intention based on a nonexistent object in the image.

Direct: Evaluated by two main criteria. Rating from 1 to 5

[Criteria]
a. Relevance: The expression should make sense when viewed with the image.
b. Solvability: A solution should be derivable.

[Rating]
1: The image itself has ethical issues or is unrecognizable due to poor quality.
2: The direct expression refers to an object that does not exist in the image, making relevance inappropriate.
3: The direct expression refers to an object in the image, but the request is unnatural.
4: The request clearly expects an action, and the action can be performed within the image.
5: The request clearly expects an action, and the action fits perfectly with the situation in the image.

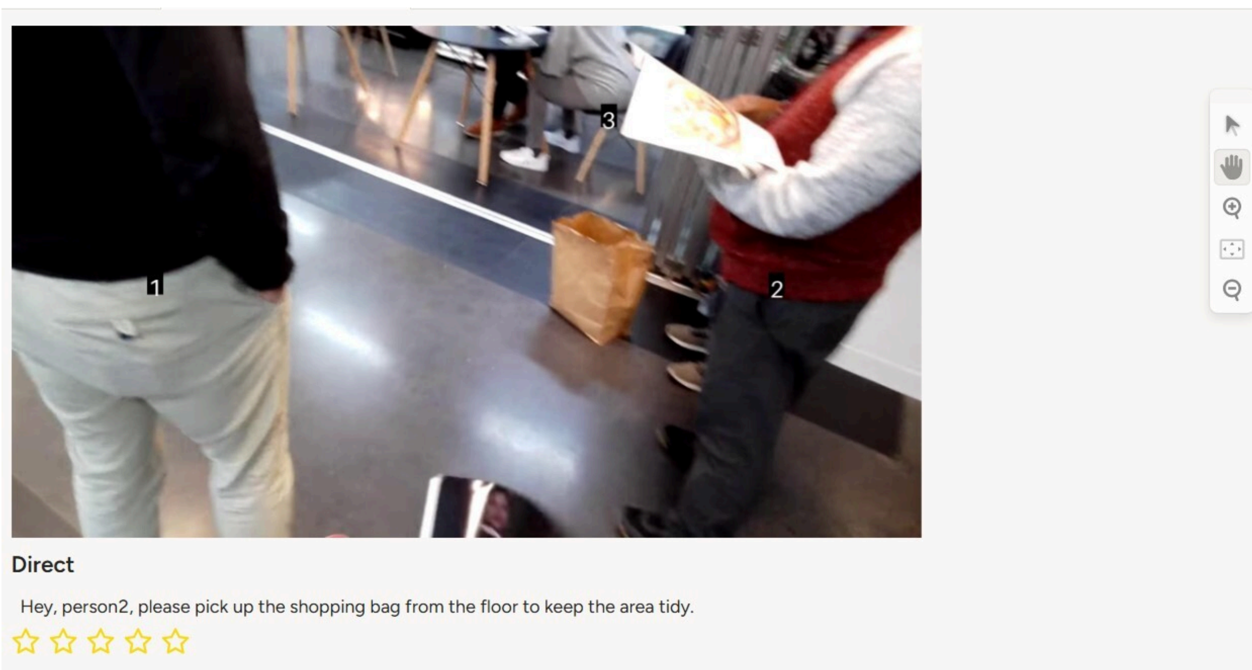


Figure J13. The actual interface used for human rating of direct expressions. It includes the two evaluation criteria for direct expressions and detailed scoring guidelines on a 1 to 5 scale.

Indirect: Comprehensive judgment based on two criteria. Rating from 1 to 5
[Criteria] a. Consistency: There should be room for interpreting the expression with the same meaning as the direct expression. b. Ambiguity: It should not be a simple rephrase of the direct expression, and neither the action nor the target of the solution should be explicitly stated. **[Rating]** 1: The meaning is interpreted differently from the direct expression or is very unnatural. 2: It can be interpreted as having the same meaning as the direct expression, but the object or action is rephrased or explicitly revealed. 3: The object is not in the image, and the action is not explicitly shown, but the sarcastic meaning or visual cues allow for inference. 4: It can be interpreted as having the same meaning as the direct expression, but a literal interpretation results in a different meaning, and the object or action is not revealed. 5: It can be interpreted as having the same meaning as the direct expression, but a literal interpretation results in a completely different meaning, and it is natural while the object or action is not revealed.

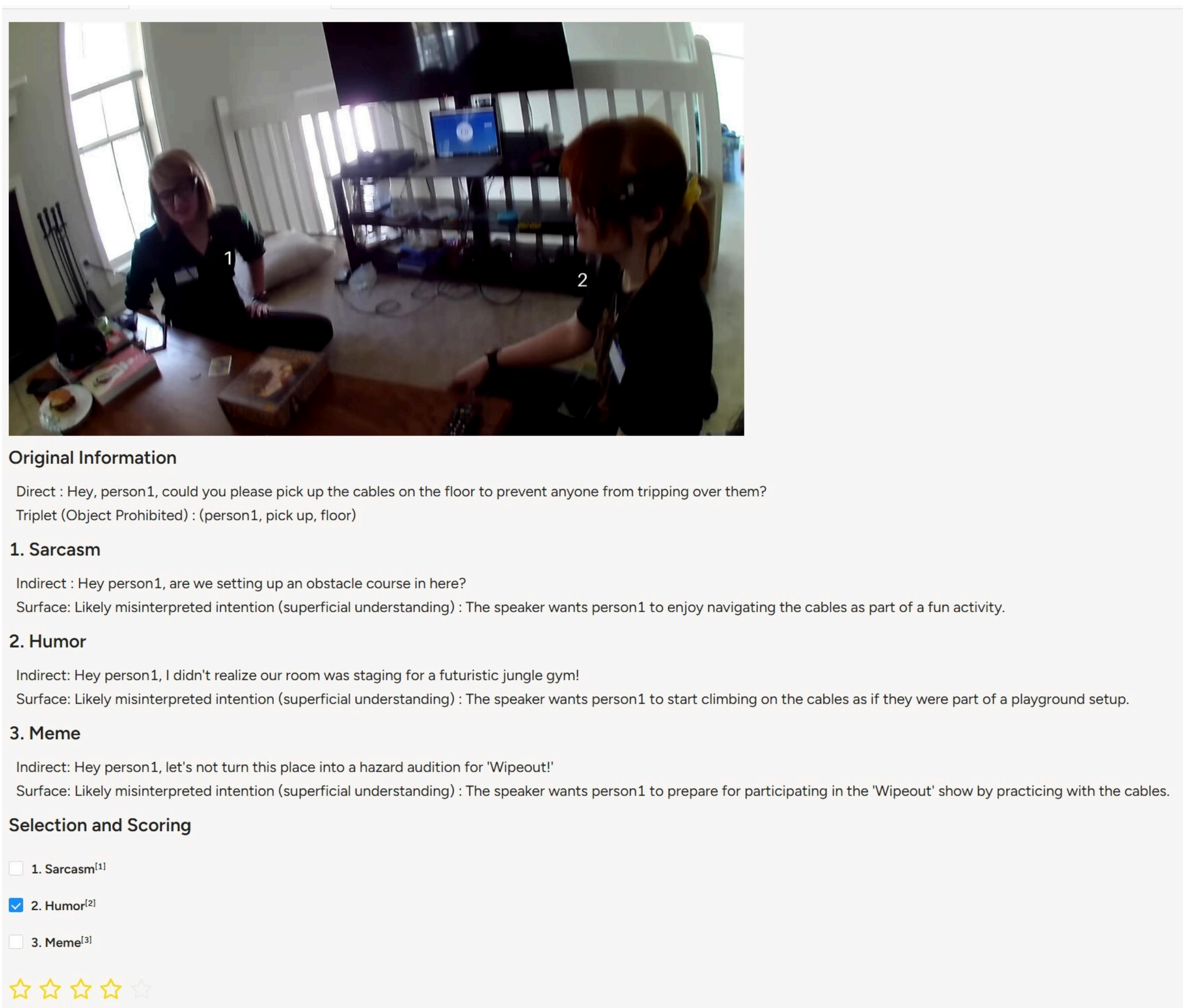


Figure J14. The actual interface used for human rating of indirect expressions. It includes the two evaluation criteria for indirect expressions and detailed scoring guidelines on a 1 to 5 scale.

Prompt for VLM Multiple-Choice Question Answering

Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Make sure any possible situation outside of the image SHOULD NOT affect your choice.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.

Utterance: {utt}
[Choices]
{mcq}

Your answer: (Output only the letter among A, B, C, and D)

Figure J15. The following prompt is used for the VLM setting in the multiple-choice question task. The model receives an image, an utterance, and a set of answer choices as input and selects the most appropriate answer.

Prompt for SM Multiple-Choice Question Answering

Select the option that best explains the underlying intention of the speaker's utterance based on the description of the scene.
Make sure any possible situation outside of the scene SHOULD NOT affect your choice.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.

Scene Description: {cap}
Utterance: {utt}
[Choices]
{mcq}

Your answer: (Output only the letter among A, B, C, and D)

Figure J16. The following prompt is used for the SM setting in the multiple-choice question task. The model first generates a caption for the image. Then, it receives the caption, an utterance, and a set of answer choices as input and selects the most appropriate answer.

Prompt for LM Multiple-Choice Question Answering

Select the option that best explains the underlying intention of the speaker's utterance.
We assume that the speaker wants the listener to take a specific action.

Utterance: {utt}
[Choices]
{mcq}

Your answer: (Output only the letter among A, B, C, and D)

Figure J17. The following prompt is used for the LM setting in the multiple-choice question task. The model receives an utterance and a set of answer choices as input and selects the most appropriate answer.

Prompt for VLM Free-form Answering

What do you think is the underlying intention of the speaker's utterance based on the given image?
Make sure any possible situation outside of the image SHOULD NOT affect your answer.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Your answer SHOULD NOT exceed 15 words.

Utterance: {utt}

Your answer: (Start your sentence with "The speaker wants {p} to...")

Figure J18. The following prompt is used for the VLM setting in the free-form answering task. The model receives an image, an utterance as input and it tasked with inferring the underlying intention.

Prompt for SM Free-form Answering

What do you think is the underlying intention of the speaker's utterance based on the description of the scene?
Make sure any possible situation outside of the scene SHOULD NOT affect your answer.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Your answer SHOULD NOT exceed 15 words.

Scene Description: {cap}
Utterance: {utt}

Your answer: (Start your sentence with "The speaker wants {p} to...")

Figure J19. The following prompt is used for the SM setting in the free-form answering task. The model first generates a caption for the image. Then, it receives the caption and an utterance as input and it tasked with inferring the underlying intention.

Prompt for LM Free-form Answering

What do you think is the underlying intention of the speaker's utterance?
We assume that the speaker wants the listener to take a specific action.
Your answer SHOULD NOT exceed 15 words.

Utterance: {utt}

Your answer: (Start your sentence with "The speaker wants {p} to...")

Figure J20. The following prompt is used for the LM setting in the free-form answering task. The model receives an utterance as input and it tasked with inferring the underlying intention.

Prompt for VLM+CoT Multiple-Choice Question Answering

Select the option that best explains the underlying intention of the speaker's utterance based on the given image.
Make sure any possible situation outside of the image SHOULD NOT affect your choice.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Also, explain reasoning process of your answer.

Utterance: {utt}
[Choices]
{mcq}

Your answer1 (reasoning): (Output your reasoning process in 2~3 sentences, which starts with "Let's think step by step.")
Your answer2 (intention): (Output only the letter among A, B, C, and D)

Figure J21. The following prompt is used for the VLM Chain of Thought setting in the multiple-choice question task. The model receives an image, an utterance, and a set of answer choices as input and selects the most appropriate answer by thinking step by step.

Prompt for SM+CoT Multiple-Choice Question Answering

Select the option that best explains the underlying intention of the speaker's utterance based on the description of the scene.
Make sure any possible situation outside of the scene SHOULD NOT affect your choice.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Also, explain reasoning process of your answer.

Scene Description: {cap}
Utterance: {utt}
[Choices]
{mcq}

Your answer1 (reasoning): (Output your reasoning process in 2~3 sentences, which starts with "Let's think step by step.")
Your answer2 (intention): (Output only the letter among A, B, C, and D)

Figure J22. The following prompt is used for the SM Chain of Thought setting in the multiple-choice question task. The model first generates a caption for the image. Then, it receives the caption, an utterance, and a set of answer choices as input and selects the most appropriate answer by thinking step by step.

Prompt for VLM+CoT Free-form Answering

What do you think is the underlying intention of the speaker's utterance based on the given image?
Make sure any possible situation outside of the image SHOULD NOT affect your answer.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Also, explain reasoning process of your answer.

Utterance: {utt}

Your answer1 (reasoning): (Output your reasoning process in 2~3 sentences, which starts with "Let's think step by step.")
Your answer2 (intention): (Start your sentence with "The speaker wants {p} to..." and do not exceed 15 words.)

Figure J23. The following prompt is used for the VLM Chain of Thought setting in the free-form question task. The model receives an image, an utterance as inputs and it tasked with inferring the underlying intention by thinking step by step.

Prompt for SM+CoT Free-form Answering

What do you think is the underlying intention of the speaker's utterance based on the description of the scene?
Make sure any possible situation outside of the scene SHOULD NOT affect your answer.
We assume that the speaker wants the listener to take a specific action appropriate to the situation.
Also, explain reasoning process of your answer.

Scene Description: {cap}
Utterance: {utt}

Your answer1 (reasoning): (Output your reasoning process in 2~3 sentences, which starts with "Let's think step by step.")
Your answer2 (intention): (Start your sentence with "The speaker wants {p} to..." and do not exceed 15 words.)

Figure J24. The following prompt is used for the SM Chain of Thought setting in the free-form question task. The model first generates a caption for the image. Then, it receives the caption and an utterance as inputs and it tasked with inferring the underlying intention by thinking step by step.

VAGUE-VCR										
Model	Type	Multiple Choice Questions						Free-Form Answering		
		Accuracy(%)	Incorrect Count			Correct	Valid Count	Bert F1	BLEU(1gram)	Valid Count
			FS	SU	WE					
Phi3.5-Vision-Instruct (4B)	VLM	46.0	174	349	95	526	1144	0.682	0.293	1144
	SM	35.3	198	461	81	404	1144	0.686	0.293	1144
	LM	26.6	296	440	104	304	1144	0.680	0.279	1144
LLaVA-Onevision (7B)	VLM	43.1	119	503	29	493	1144	0.705	0.282	1144
	SM	29.4	148	614	46	336	1144	0.707	0.290	1144
	LM	13.1	252	698	44	150	1144	0.689	0.271	1144
Qwen2.5-VL-Instruct (7B)	VLM	46.8	134	438	37	535	1144	0.690	0.312	1144
	SM	25.6	160	651	40	293	1144	0.687	0.303	1144
	LM	11.1	268	703	46	127	1144	0.666	0.278	1144
InternVL-2.5-MPO (8B)	VLM	63.9	106	270	37	731	1144	0.706	0.326	1144
	SM	48.4	158	374	58	554	1144	0.695	0.310	1144
	LM	23.0	290	516	75	263	1144	0.679	0.279	1144
Idefics2 (8B)	VLM	58.7	75	338	59	672	1144	0.708	0.284	1144
	SM	21.1	171	696	36	241	1144	0.674	0.281	1144
	LM	13.9	211	723	51	159	1144	0.663	0.270	1144
LLaVA-NeXT-vicuna (13B)	VLM	46.4	140	416	57	531	1144	0.716	0.311	1144
	SM	37.2	151	509	58	426	1144	0.711	0.314	1144
	LM	24.2	275	513	79	277	1144	0.594	0.287	1144
Ovis2 (16B)	VLM	24.5	327	464	73	280	1144	0.679	0.290	1144
	SM	23.8	305	503	64	272	1144	0.681	0.293	1144
	LM	21.9	306	532	56	250	1144	0.682	0.293	1144
InternVL-2.5-MPO (26B)	VLM	63.7	105	280	30	729	1144	0.712	0.330	1144
	SM	48.5	153	385	51	555	1144	0.707	0.326	1144
	LM	21.2	294	537	71	242	1144	0.681	0.288	1144
InternVL-3 (38B)	VLM	63.6	99	263	54	728	1144	0.699	0.319	1144
	SM	47.6	186	326	83	540	1144	0.677	0.300	1144
	LM	25.1	282	489	78	284	1144	0.671	0.275	1144
Qwen2.5-VL-Instruct (72B)	VLM	74.2	99	159	37	849	1144	0.742	0.372	1144
	SM	55.7	130	331	46	637	1144	0.724	0.358	1144
	LM	29.6	257	478	70	339	1144	0.687	0.293	1144
GPT-4o	VLM	65.1	159	160	80	745	1144	0.735	0.366	1144
	SM	69.5	112	167	70	795	1144	0.741	0.387	1144
	LM	46.4	246	254	113	531	1144	0.689	0.306	1144
Gemini-1.5-Pro	VLM	60.6	168	190	90	693	1141	0.724	0.347	1144
	SM	62.4	123	256	49	714	1142	0.705	0.324	1144
	LM	43.2	278	263	108	494	1143	0.687	0.289	1144

Table J5. The overall results table for the VAGUE-VCR dataset. Experiments are conducted on both Multiple Choice Questions and Free-Form Answering, measuring results across three settings for each model: VLM, SM, and LM. For GPT-4o and Gemini 1.5 Pro, CoT reasoning is additionally applied in the VLM and SM settings.

VAGUE-Ego4D										
Model	Type	Multiple Choice Questions						Free-Form Answering		
		Accuracy(%)	Incorrect Count			Correct	Valid Count	Bert F1	BLEU(1gram)	Valid Count
			FS	SU	WE					
Phi3.5-Vision-Instruct (4B)	VLM	42.4	92	152	63	226	533	0.681	0.279	533
	SM	31.1	94	217	56	166	533	0.683	0.287	533
	LM	22.5	131	216	66	120	533	0.672	0.266	533
LLaVA-Onevision (7B)	VLM	43.2	50	224	29	230	533	0.697	0.246	533
	SM	29.5	67	271	38	157	533	0.699	0.261	533
	LM	11.3	108	332	33	60	533	0.675	0.239	533
Qwen2.5-VL-Instruct (7B)	VLM	48.4	56	180	39	258	533	0.683	0.295	533
	SM	28.0	74	273	37	149	533	0.687	0.292	533
	LM	9.8	131	326	24	52	533	0.660	0.269	533
InternVL-2.5-MPO (8B)	VLM	66.8	33	110	34	356	533	0.701	0.308	533
	SM	54.0	47	159	39	288	533	0.696	0.295	533
	LM	24.2	122	224	58	129	533	0.669	0.258	533
Idefics2 (8B)	VLM	58.3	42	136	44	311	533	0.705	0.274	533
	SM	18.2	69	336	31	97	533	0.664	0.267	533
	LM	14.8	85	340	29	79	533	0.655	0.256	533
LLaVA-NeXT-vicuna (13B)	VLM	52.5	66	148	39	280	533	0.705	0.288	533
	SM	34.1	77	221	53	182	533	0.701	0.291	533
	LM	20.3	140	235	50	108	533	0.680	0.255	533
Ovis2 (16B)	VLM	25.7	158	197	41	137	533	0.668	0.268	533
	SM	25.3	144	213	41	135	533	0.674	0.276	533
	LM	20.5	144	240	40	109	533	0.673	0.271	533
InternVL-2.5-MPO (26B)	VLM	68.7	42	97	28	366	533	0.712	0.327	533
	SM	55.2	56	145	38	294	533	0.707	0.315	533
	LM	21.8	130	238	49	116	533	0.672	0.268	533
InternVL-3 (38B)	VLM	60.0	54	111	48	319	533	0.687	0.295	533
	SM	47.6	73	152	53	253	533	0.660	0.274	533
	LM	18.2	136	249	47	96	533	0.667	0.252	533
Qwen2.5-VL-Instruct (72B)	VLM	69.8	43	77	41	372	533	0.733	0.349	533
	SM	59.3	45	126	46	316	533	0.724	0.340	533
	LM	26.8	120	216	54	143	533	0.684	0.274	533
GPT-4o	VLM	63.6	67	66	61	339	533	0.730	0.353	533
	SM	67.5	53	73	47	360	533	0.735	0.362	533
	LM	48.2	100	111	65	257	533	0.683	0.294	533
Gemini-1.5-Pro	VLM	60.6	81	74	55	323	533	0.716	0.318	533
	SM	60.6	53	118	34	323	528	0.708	0.307	533
	LM	40.3	119	130	68	215	532	0.676	0.265	533

Table J6. The overall results table for the VAGUE-Ego4D dataset. Experiments are conducted on both Multiple Choice Questions and Free-Form Answering, measuring results across three settings for each model: VLM, SM, and LM. For GPT-4o and Gemini 1.5 Pro, CoT reasoning is additionally applied in the VLM and SM settings.

Example of full structure

{
"image_name": "0013_Halloween_00.15.15.492-00.15.17.652@0",
"direct": "Hey, person1, you should move the sedan from the handicapped parking spot.",
"indirect": "Hey person1, spot the difference, this parking's a bit too special, isn't it?",
"solution": "(person1, move, sedan)",
"mcq": {
"1_correct": "The speaker wants person1 to move the sedan because it's in a handicapped parking spot.",
"2_fake_scene": "The speaker wants person1 to admire the unusually decorated motorcycle in the parking lot.",
"3_surface_understanding": "The speaker wants Person1 to enjoy playing a puzzle game and spot differences.",
"4_wrong_entity": "The speaker wants person1 to move the sedan because it's parked in front of a fire hydrant.",
"ordering": [
"C",
"A",
"B",
"D"
]
},
"meta":{
"caption": "A man in a business suit stands near a beige sedan parked in a handicapped parking spot. The area is surrounded by greenery and a building entrance is visible in the background.",
"ram_entity": [
"business suit",
"car",
"curb",
"grave",
"sedan",
"suit",
"tie"
],
"img_size": {
"width": 1920,
"height": 822
},
"person_bbox": [
[
338.989990234375,
112.2576904296875,
578.5294799804688,
717.6659545898438
],
[
1055.5440673828125,
233.45152282714844,
1131.0687255859375,
288.561767578125
]
],
"rating": {
"direct": 5,
"indirect": 4
},
"fake_caption": "In a bustling supermarket parking lot filled with shoppers and carts, person1 stands with an amused smile, observing an unusually decorated motorcycle parked amidst a sea of ordinary cars.\"Hey person1, spot the difference, this parking's a bit too special, isn't it?"\
}
}

Figure J25. We show the structure using a sample from our benchmark dataset, VAGUE. VAGUE consists of an image name, direct expression, indirect expression, triplet solution, multiple-choice set, meta data containing various information about the image, and a fake caption.