PhD: A ChatGPT-Prompted Visual hallucination Evaluation Dataset

Supplementary Material

In the supplementary materials, we report

- Manual assessment of the dataset quality (Sec. S1);
- ChatGPT instructions for dataset construction (Sec. S2);
- Additional experimental results (Sec. S3);
- `PhD` question-answer data (`data.json`) (Sec. S4).

S1. Manual Assessment of Dataset Quality

We assessed the quality of `PhD` by randomly selecting 100 samples (*i.e. image-question-answer* triplets) per mode / task. Each sample has a specific image associated with a `Yes` question and a `No` question. As an image might have multiple questions while a sample is shared across modes and tasks, the subset has 885 distinct images and 1,794 samples in total, see Tab. S1. Manual assessment was conducted with the following evaluation criteria:

- The VQA data must be correct, with a correct `Yes` / `No` label w.r.t. the given image and question;
- The question (and the preceded context) shall effectively induce hallucinations, with the hitem, context, and CCS images designed to be challenging.

Samples meeting both criteria are labeled as `good`, and `bad` otherwise. Our annotation team was comprised of 9 volunteers (4 phd and 5 graduate students in our lab), who have good knowledge about computer vision and deep learning, yet do not involve in the current project.

The labeling result is analyzed and summarized in Tab. S1, showing that the sampled subset of `PhD` achieves a good rate of 93%. Note that the annotation process has been conducted in a collective manner where a specific sample was assessed independently by different annotators, labeling discrepancy naturally exists. The relatively small deviation of 2% on average w.r.t. the `good` rate indicates a consensus of opinions among the annotators.

	#images	#VQA Triplets	#good	good-rate
Mode				
PhD-base	100	200	188	94%±2%
PhD-sec	100	200	185	93%±3%
PhD-icc	100	200	192	96%±2%
PhD-ccs	100	200	192	96%±1%
Task				
Object recog.	100	200	189	95%±2%
Attribute recog.	100	200	188	94%±2%
Sentiment recog.	100	200	179	90%±3%
Positional recog.	100	200	178	89%±4%
Counting	100	200	184	92%±3%
Overall	885	1,794	1,675	93%±2%

Table S1. **Dataset quality assessment**. #good indicates the number of sampled triplets manually assessed as `good`.

S2. ChatGPT Instructions for Dataset Construction

ChatGPT instructions used in each step of the proposed pipeline are listed as follows.

S2.1. Instructions for Vocabulary Construction

```
Targeted word dictionary is like:
 {
   "object": ['car', 'bus', 'person' ...],
   "attribute_color": ['red', 'blue',
       ...],
   "attribute_material": ['plastic', ...],
   "attribute_shape": ['square', ...],
   "sentiment": ['happy', ...]
   # "counting": [...] zero to nine
   # "position": [...] similar to object
 }

 Please expand the input dictionary as
     much as possible by adding common
     items found in daily life under each
     key. Avoid any duplication.
 Output the result directly in <Python
     dictionary> format without providing
     any explanation or comments.

 output:
```

Prompt S1. Vocabulary construction

S2.2. Instructions for Subject-Attribute Extraction from TDIUC Annotations

```
Based on the provided question <q> and <
    task>, extract the specific subject
    being asked about in the question.
Output requirement: please output only the
    subject without any explanation or
    additional information.

Here are some examples:
task: positional_reasoning
q: What is to the right of the book?
output: to the right of the book

task: attribute_color
q: What is the color of the grass?
output: grass

task: attribute_material
q: What is the chair made of?
```

```
output: chair

task: sentiment_understanding
q: Is this man happy?
output: man

task: counting
q: How many people are there?
output: people

Now, following the examples above, please extract the relevant subject from the question without providing any explanation or comments.

task: %s
q: %s
output:
```

Prompt S2. Subject-attribute extraction

S2.3. Instructions for Hitem-embedded question generation

```
 Please construct a binary question based on the following input, requiring the question to be answered with "yes" or "no" and clearly highlighting the task. Ensure the question is concise, grammatically correct, and clearly understood.

For example:

input: red
related object: apple
task: attribute_color
output: Is the apple in the image red?

input: airplane
related object: None
task: object
output: Is there an airplane in the image?

input: circle
related object: plate
task: attribute_shape
output: Is the plate circular in shape?

input: plastic
related object: bottle
task: attribute_material
output: Is the bottle made of plastic in the image?

input: position_reasoning
related object: person
task: to the left of the tv
output: Is there a person to the left of the TV in the image?

input: happy
related object: dog
task: sentiment_understanding
output: Is the dog happy in the image?

input: 3
related object: bags
task: counting
output: Is the number of the bags in the image 3?

Please learn from the above example and construct a yes/no question based on the following input, related object, and task. Do not provide any explanations or clarifications.

input: %s
related object: %s
task: %s
output:
```

Prompt S3. Hitem-embedded question generation

S2.4. Instructions for Specious Text Generation

```
Please generate the <specious text> for the given question. It should be one sentence.

The <specious text> should answer the question, but it may not reflect the actual current status, thus making it specious. Avoid using words like "Image" or "Photo."

Using status and time words like 'all the time,' 'from time to time,' 'always,' and 'usually' and so on can greatly help with this task.

Examples:

Captions: ['Some parents sitting behind a fence watching a baseball game', 'Group of parents watching small children on a baseball field.', ...]
Question: Is the chair to the right of the person in the image?
Answer: Yes
specious text: Seasoned spectators know that the best spot of chairs is always to the right side, providing a clear view of all the action.

```

```
Captions: [COCO captions]
Question: Is there a piece of butter in the image?
Answer: No
Unverifiable context: They never use butter in their cake, so they do not prepare any on the wooden table.

Captions: [COCO captions]
Question: Is there a piece of butter in the image?
Answer: Yes
specious text: The croissants and cake are usually made even more delicious with a certain butter addition.

Captions: [COCO captions]
Question: Is there a bird in the image?
Answer: Yes
specious text: The water is not only cool, but also countless unknown water birds can be seen from time to time.

Captions: [COCO captions]
Question: Is the plate oval in shape?
Answer: no
Unverifiable context: The square plate is usually the perfect choice for dessert, adding a touch of elegance to the presentation.

Captions: [COCO captions]
Question: Are there 3 people in the image?
Answer: yes
specious text: Three of the audience members stood up excitedly all the time, ready to cheer on their favorite player.

Captions: [COCO captions]
Question: Is the plate oval in shape?
Answer: yes
specious text: The oval plate is usually the perfect choice for dessert, adding a touch of elegance to the presentation.

Captions: captions
Question: Is there a cup in the image?
Answer: no
Unverifiable context: This table is used for food, thus the cups are usually placed on the other tables.

Captions: captions
Question: Is there a ball in the image?
Answer: No
Unverifiable context: The ball has been gone recently. Without the companionship of a ball, the kitten just finds a place to rest and sleep.


Captions: [COCO captions]
Question: Are there 3 sandwiches in the image?
Answer: yes
specious text: Ordering three pizzas at once often comes with significant discounts, so people frequently choose this option.

Generate the specious text directly without any explanations.

Captions: %s
Question: %s
Answer: %s
specious text:
```

Prompt S4. Specious text generation

S2.5. Instructions for (Specious / Incorrect Context) Text Composition

```
Please combine the <input> and <captions> into a complete context that reads like a web article or news report.
1. Ensure that the <input> content is prioritized and fully integrated, while the <captions> serve to provide background information.
2. Avoid using words like "Image" or "Photo" and ensure the final context does not resemble an image description.
3. The final context should be natural and smooth.
4. When the <input> conflicts with the <captions>, trust the <input>.

Please directly output the combined context without any explanation or clarification. The output should be concise, consisting of 2 to 3 sentences.

captions: %s
input: %s
output:
```

Prompt S5. Text composition (specious context)

S2.6. Instructions for CCS Description Generation

```
Please generate abnormal visual content and its corresponding normal content based on the specified type. There are five
```

```
    types of visual tasks you can generate
    content.

Examples:

task: object
abnormal: Ice blocks in volcanic lava
normal: Fire in volcanic lava
task: object
abnormal: Milk flowing from a faucet
normal: Water flowing from a faucet
task: object
abnormal: A table in an intersection
normal: Cars in an intersection
task: object
abnormal: Grass in the tiger's mouth
normal: Meat in the tiger's mouth

task: attribute
abnormal: A car with square wheels
normal: A car with round wheels
task: attribute
abnormal: Blue apples on the tree
normal: Red apples on the tree
task: attribute
abnormal: A purple sky
normal: A blue sky

task: count
abnormal: A die with a maximum of 7 dots
normal: A die with a maximum of 6 dots
task: count
abnormal: A person with 7 fingers on one
    hand
normal content: A person with 5 fingers on
    one hand
task: count
abnormal: A table with 3 legs
normal: A table with 4 legs

task: sentiment
abnormal: A rabbit smiling in front of a
    tiger
normal: A rabbit fearful in front of a
    tiger
task: sentiment
abnormal: A person is relaxed at a busy
    intersection
normal: A person hurrying at a busy
    intersection
task: sentiment

task: position
abnormal: A tree with roots at the top and
    leaves at the bottom
normal: A tree with roots at the bottom and
     leaves at the top
task: position
abnormal: A television inside a fish tank
normal: A television on a stand

Please learn from the examples above and
    generate a new pair of abnormal and
    normal content based on the specified
    type. Use the given format and do not
    provide any additional explanations or
    comments.

type: %s
```

Prompt S6. CCS description generation

S2.7. Instructions for Question Generation from CCS Descriptions

```
Please Convert the description into a
    binary question format.

Examples:
description: the car with square wheels.
question: Does the car have square wheels?

description: books in a swimming pool
question: Are there books in a swimming
    pool?

Please learn from the examples above to
    generate questions. Respond only with
    the binary question format, without
    explanations or comments.

description: %s
question:
```

Prompt S7. Question Generation

S3. Experimental Details and Extra Results

S3.1. MLLM-specific Prompts

The following are official prompts for each MLLM, as mentioned in Sec. 4.1 (*Test Protocol*).

```
<s>[INST] <Img><ImageHere></Img> [vqa]{
    PhD_question} [/INST]
```

Prompt S8. Evaluation prompt for MiniGPT-v2

```
A chat between a curious human and an
    artificial intelligence assistant. The
    assistant gives helpful, detailed, and
    polite answers to the human's questions.
     USER: <image>
{PhD_question} ASSISTANT:
```

Prompt S9. Evaluation prompt for LLaVA-1.5

```
USER: <|image|>{PhD_question} ASSISTANT:
```

Prompt S10. Evaluation prompt for mPLUG-Owl2

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
Picture 1:<img></img>
{PhD_question}<|im_end|>
<|im_start|>assistant
```

Prompt S11. Evaluation prompt for Qwen-VL

```
<ImageHere>{PhD_question}
```

Prompt S12. Evaluation prompt for InstructBLIP

```
A chat between a curious human and an
    artificial intelligence assistant. The
    assistant gives helpful, detailed, and
    polite answers to the human's questions.
     USER: <image>
{PhD_question} ASSISTANT:
```

Prompt S13. Evaluation prompt for LLaVA-1.6

```
A chat between a curious human and an
    artificial intelligence assistant. The
    assistant gives helpful, detailed, and
    polite answers to the human's questions.
     USER: <image>
{PhD_question} ASSISTANT:
```

Prompt S14. Evaluation prompt for LLaVA-1.1

`PhD_question` means specific questions provided in `PhD`. During evaluations on `PhD`-sec and `PhD`-icc, the context is prepended to `PhD_question`, with the instruction *"If the context conflicts with the image, prioritize the image."* inserted in between.

S3.2. Detailed Experimental Results

We provide the detailed performance results of the experiments discussed in Sec. 4, as shown in Tab. S2 and Tab. S3.

Mode-Oriented VHE performance. Tab. S2 demonstrate the specific performance corresponding to Fig. 4.a. GPT-4o outperforms open-source models across all modes, achieving a more balanced performance between yes-recall and no-recall.

Task-Oriented VHE performance. Tab. S3 demonstrate the specific performance corresponding to Fig. 4.b. The five visual tasks form a difficulty hierarchy: as the required feature granularity increases, the model's performance declines. Notably, GPT-4o exhibits slightly weaker performance in sentiment recognition, likely due to the more open-ended nature of this task. Emotions such as happiness, joy, and excitement can be interpreted as similar, making finer distinctions potentially detrimental to accurate identification.

S3.3. Additional Experiments

Comparison with AMBER. As previously analyzed, AMBER's reliance on manual annotation leads to the selection of overly simple images to maintain annotation accuracy, as illustrated in Fig. S1, Fig. S2, Fig. S3, Fig. S4. Using DETR [S1], we observed that the average number of objects detected in AMBER images is 2.3, significantly lower than the 10.6 objects detected in images used in `PhD`. Furthermore, the question format and content in AMBER are simplistic, whereas `PhD`offers significantly richer content. The images in `PhD`are more complex and closer to daily scenarios, making it more challenging and a valuable contribution to existing objective evaluation benchmarks.

Consistency matrix among MLLMs. In Fig. S5, we present consistency matrices of selected MLLMs across four modes. Generally, models from the same series exhibit higher correlations. For instance, in `PhD`-base, `PhD`-sec, and `PhD`-icc, LLaVA-1.5, LLaVA-1.5-L, LLaVA-1.6, and LLaVA-1.6-L show strong correlations. However, in `PhD`-ccs, the correlations among all models are relatively low. This is because `PhD`-ccs evaluates more complex hallucination causes, which are influenced by factors such as model size, visual capabilities, and training strategies. As a result, even models from the same series exhibit varying behaviors in this mode.

CLIP scores on `PhD`-ccs. We calculate the CLIP scores between CS/CCS descriptions and CCS images. The mean cosine similarity between CS descriptions and mages is 0.20, while CCS descriptions achieve 0.25. This indicates that CLIP itself can detect conter-common-sense content, as the visual elements involved are simple (*e.g.*, the sizes of a mouse and a cat). Nonetheless, the low performance on `PhD`-ccs demonstrates that MLLMs tend to ignore such counterintuitive visual cues and instead rely on their internal knowledge.

S4. Annotation Files of `PhD`

In the supplementary materials, we also provide the annotation file *data.json* for peer review. Note that due to upload size limitations, the corresponding images are not included. The description and use of the annotation files can be found in the accompanying *README.md* file.

References

[S1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 5

MLLM	Visual encoder	LLM kernel	PhD-base			PhD-sec			PhD-icc			PhD-ccs			PhD Index
			Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	
GPT-4o	–	–	0.83± 0.09	0.86± 0.09	0.84± 0.08	0.80± 0.11	0.86± 0.11	0.83± 0.10	0.70± 0.13	0.87± 0.13	0.77± 0.12	0.75± 0.09	0.84± 0.09	0.79± 0.09	0.81
Claude 3.5 Sonnet	–	–	0.75± 0.11	0.78± 0.11	0.76± 0.10	0.72± 0.08	0.88± 0.08	0.79± 0.09	0.67± 0.12	0.85± 0.12	0.73± 0.15	0.66± 0.12	0.77± 0.12	0.71± 0.08	0.75
Gemini 1.5 Pro	–	–	0.81± 0.08	0.85± 0.08	0.83± 0.10	0.64± 0.07	0.88± 0.07	0.73± 0.12	0.50± 0.12	0.74± 0.12	0.56± 0.24	0.55± 0.11	0.79± 0.11	0.64± 0.05	0.69
Qwen-VL (Woodpecker)	-bigG/14	Qwen-7B	0.63± 0.12	0.59± 0.12	0.59± 0.07	0.61± 0.20	0.54± 0.20	0.53± 0.10	0.58± 0.06	0.78± 0.06	0.65± 0.09	0.28± 0.12	0.64± 0.08	0.39± 0.11	0.53
LLaVA-1.6-L (Woodpecker)	-L/14	Vicuna-13B-1.5	0.59± 0.30	0.46± 0.30	0.44± 0.29	0.41± 0.09	0.82± 0.09	0.52± 0.14	0.40± 0.31	0.71± 0.31	0.44± 0.23	0.17± 0.09	0.80± 0.11	0.29± 0.12	0.41
LLaVA-OneVision	SoViT-400m/14	Qwen2-72B	0.84± 0.09	0.73± 0.09	0.78± 0.07	0.76± 0.11	0.64± 0.11	0.69± 0.09	0.69± 0.13	0.59± 0.13	0.63± 0.12	0.69± 0.11	0.71± 0.11	0.70± 0.07	0.70
Molmo	-L/14	Qwen2-72B	0.84± 0.13	0.65± 0.13	0.73± 0.11	0.62± 0.10	0.82± 0.10	0.70± 0.11	0.51± 0.09	0.78± 0.09	0.60± 0.16	0.67± 0.07	0.84± 0.07	0.74± 0.04	0.69
InternVL-1.5	InternViT-6B	InternLM2-20B	0.89± 0.13	0.66± 0.13	0.75± 0.09	0.55± 0.11	0.64± 0.11	0.59± 0.15	0.39± 0.11	0.31± 0.11	0.30± 0.16	0.48± 0.10	0.80± 0.10	0.60± 0.09	0.56
Qwen-VL (VCD)	-bigG/14	Qwen-7B	0.81± 0.08	0.71± 0.08	0.76± 0.07	0.59± 0.16	0.50± 0.16	0.50± 0.14	0.52± 0.09	0.22± 0.09	0.28± 0.12	0.74± 0.12	0.68± 0.12	0.71± 0.09	0.56
Cambrian-1	Hybrid	Llama-3-8B	0.85± 0.11	0.66± 0.11	0.74± 0.09	0.65± 0.09	0.44± 0.09	0.52± 0.10	0.55± 0.12	0.24± 0.12	0.30± 0.14	0.60± 0.14	0.68± 0.14	0.63± 0.10	0.55
LLaVA-1.6-L (VCD)	-L/14	Vicuna-13B-1.5	0.81± 0.08	0.59± 0.08	0.68± 0.07	0.63± 0.08	0.34± 0.08	0.44± 0.09	0.55± 0.05	0.20± 0.05	0.29± 0.07	0.71± 0.10	0.58± 0.10	0.64± 0.08	0.51
LLaVA-1.6-XL	-L/14	Nous-Hermes-2-Yi-34B	0.78± 0.12	0.57± 0.12	0.65± 0.09	0.70± 0.06	0.34± 0.06	0.46± 0.07	0.54± 0.03	0.16± 0.03	0.23± 0.02	0.72± 0.16	0.57± 0.16	0.62± 0.13	0.49
Qwen-VL	ViT-bigG/14	Qwen-7B	0.88± 0.11	0.56± 0.11	0.68± 0.09	0.40± 0.17	0.48± 0.17	0.42± 0.14	0.12± 0.10	0.25± 0.10	0.16± 0.06	0.72± 0.16	0.67± 0.16	0.69± 0.12	0.49
LLaVA-1.6-L	ViT-L/14	Vicuna-13B-1.5	0.85± 0.17	0.58± 0.17	0.65± 0.07	0.61± 0.23	0.40± 0.23	0.37± 0.09	0.26± 0.20	0.18± 0.20	0.11± 0.04	0.49± 0.15	0.71± 0.15	0.56± 0.04	0.42
MiniGPT-v2	ViT-G/14	LLaMA-2-7B	0.76± 0.11	0.68± 0.11	0.71± 0.08	0.45± 0.06	0.16± 0.06	0.23± 0.07	0.21± 0.02	0.03± 0.02	0.05± 0.03	0.69± 0.07	0.48± 0.07	0.57± 0.08	0.39
LLaVA-1.6	ViT-L/14	Vicuna-7B-1.5	0.85± 0.18	0.57± 0.18	0.64± 0.08	0.70± 0.10	0.15± 0.10	0.21± 0.09	0.36± 0.03	0.02± 0.03	0.03± 0.03	0.68± 0.20	0.63± 0.20	0.61± 0.08	0.37
MPlug-Owl2	ViT-L/14	LLaMA-2-7B	0.93± 0.08	0.34± 0.08	0.50± 0.09	0.70± 0.04	0.08± 0.04	0.14± 0.06	0.33± 0.01	0.01± 0.01	0.01± 0.01	0.88± 0.18	0.51± 0.18	0.63± 0.14	0.32
InstructBLIP	ViT-G/14	Vicuna-7B-1.1	0.89± 0.20	0.42± 0.20	0.54± 0.19	0.71± 0.04	0.06± 0.04	0.09± 0.06	0.38± 0.01	0.01± 0.01	0.01± 0.01	0.74± 0.05	0.48± 0.05	0.58± 0.06	0.31
InstructBLIP-L	ViT-G/14	Vicuna-13B-1.1	0.96± 0.14	0.21± 0.14	0.32± 0.19	0.64± 0.05	0.17± 0.05	0.26± 0.07	0.32± 0.01	0.03± 0.01	0.05± 0.02	0.83± 0.14	0.36± 0.14	0.48± 0.13	0.28
LLaVA-1.5-L	ViT-L/14	Vicuna-13B-1.5	0.96± 0.07	0.27± 0.07	0.42± 0.08	0.72± 0.02	0.05± 0.02	0.10± 0.04	0.24± 0.01	0.01± 0.01	0.02± 0.02	0.93± 0.21	0.42± 0.21	0.54± 0.21	0.27
LLaVA-1.5	ViT-L/14	Vicuna-7B-1.5	0.95± 0.10	0.30± 0.10	0.44± 0.11	0.71± 0.03	0.04± 0.03	0.08± 0.06	0.30± 0.00	0.01± 0.00	0.01± 0.01	0.90± 0.19	0.41± 0.19	0.53± 0.18	0.27
LLaVA-1.1	ViT-L/14	Vicuna-7B-1.1	0.98± 0.05	0.07± 0.05	0.13± 0.09	0.69± 0.00	0.01± 0.00	0.02± 0.01	0.27± 0.01	0.00± 0.01	0.01± 0.01	0.84± 0.15	0.26± 0.15	0.38± 0.16	0.14

Table S2. **Mode-oriented performance of varied MLLMs on `PhD`**. The best open-source model per mode is highlighted in green. Values after the symbol ± indicate the standard deviation across the five tasks.

MLLM	Object Recognition			Attribute Recognition			Sentiment Recognition			Positional Recognition			Counting		
	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index	Yes-Recall	No-Recall	PhD Index
GPT-4o	0.83±0.07	0.92±0.03	0.87±0.05	0.84±0.06	0.89±0.02	0.86±0.03	0.64±0.10	0.71±0.11	0.67±0.10	0.80±0.05	0.93±0.04	0.86±0.02	0.74±0.13	0.85±0.11	0.79±0.11
Claude 3.5 Sonnet	0.85±0.05	0.88±0.05	0.86±0.10	0.77±0.07	0.85±0.07	0.81±0.04	0.66±0.12	0.70±0.12	0.68±0.07	0.65±0.07	0.91±0.07	0.76±0.04	0.55±0.14	0.76±0.14	0.62±0.09
Gemini 1.5 Pro	0.84±0.03	0.89±0.03	0.86±0.09	0.58±0.04	0.82±0.04	0.66±0.15	0.55±0.12	0.72±0.12	0.62±0.08	0.69±0.06	0.90±0.06	0.78±0.11	0.46±0.12	0.75±0.12	0.53±0.20
Qwen-VL (Woodpecker)	0.77±0.06	0.68±0.06	0.71±0.01	0.49±0.02	0.71±0.02	0.58±0.05	0.50±0.05	0.52±0.05	0.50±0.07	0.51±0.04	0.89±0.04	0.65±0.05	0.72±0.18	0.47±0.18	0.55±0.12
LLaVA-1.6-L (Woodpecker)	0.67±0.08	0.75±0.08	0.70±0.01	0.34±0.03	0.84±0.03	0.48±0.05	0.29±0.08	0.64±0.08	0.38±0.10	0.36±0.39	0.56±0.39	0.26±0.19	0.74±0.18	0.67±0.18	0.69±0.11
LLaVA-OneVision	0.84±0.05	0.79±0.05	0.82±0.06	0.78±0.06	0.74±0.06	0.76±0.07	0.73±0.13	0.56±0.13	0.63±0.09	0.77±0.07	0.67±0.07	0.71±0.04	0.63±0.09	0.59±0.09	0.61±0.08
Molmo	0.83±0.05	0.80±0.05	0.81±0.08	0.64±0.05	0.89±0.05	0.73±0.11	0.62±0.14	0.64±0.14	0.62±0.08	0.66±0.13	0.79±0.13	0.70±0.03	0.55±0.05	0.72±0.05	0.60±0.16
InternVL-1.5	0.74±0.15	0.76±0.15	0.73±0.12	0.47±0.19	0.64±0.19	0.51±0.28	0.61±0.21	0.50±0.21	0.52±0.13	0.62±0.24	0.59±0.24	0.57±0.17	0.46±0.17	0.52±0.17	0.47±0.19
Qwen-VL (VCD)	0.84±0.17	0.63±0.17	0.71±0.12	0.65±0.25	0.55±0.25	0.59±0.23	0.63±0.28	0.48±0.28	0.52±0.24	0.78±0.21	0.46±0.21	0.55±0.20	0.43±0.16	0.52±0.16	0.44±0.19
Cambrian-1	0.78±0.20	0.63±0.20	0.66±0.15	0.54±0.21	0.56±0.21	0.54±0.23	0.64±0.13	0.52±0.13	0.57±0.10	0.74±0.22	0.49±0.22	0.55±0.19	0.60±0.18	0.34±0.18	0.41±0.19
LLaVA-1.6-L (VCD)	0.78±0.16	0.52±0.16	0.61±0.13	0.58±0.21	0.43±0.21	0.49±0.21	0.69±0.20	0.42±0.20	0.51±0.18	0.75±0.16	0.38±0.16	0.48±0.16	0.56±0.15	0.42±0.15	0.48±0.15
LLaVA-1.6-XL	0.82±0.21	0.48±0.21	0.58±0.19	0.58±0.20	0.46±0.20	0.51±0.21	0.71±0.21	0.38±0.21	0.47±0.18	0.75±0.18	0.39±0.18	0.48±0.17	0.57±0.15	0.34±0.15	0.42±0.14
Qwen-VL	0.61±0.28	0.61±0.11	0.59±0.20	0.56±0.28	0.53±0.23	0.54±0.25	0.47±0.34	0.41±0.27	0.42±0.29	0.56±0.31	0.46±0.19	0.49±0.24	0.46±0.29	0.43±0.15	0.40±0.16
LLaVA-1.6-L	0.63±0.29	0.47±0.25	0.50±0.22	0.54±0.28	0.43±0.23	0.46±0.21	0.27±0.17	0.80±0.14	0.38±0.21	0.64±0.27	0.36±0.24	0.41±0.21	0.68±0.16	0.28±0.18	0.37±0.21
MiniGPT-v2	0.64±0.23	0.40±0.31	0.46±0.29	0.49±0.28	0.36±0.29	0.40±0.30	0.40±0.25	0.32±0.27	0.35±0.27	0.59±0.24	0.29±0.23	0.37±0.27	0.53±0.10	0.33±0.20	0.38±0.20
LLaVA-1.6	0.70±0.24	0.37±0.28	0.43±0.29	0.65±0.26	0.29±0.27	0.35±0.31	0.34±0.16	0.56±0.37	0.41±0.23	0.78±0.18	0.28±0.26	0.35±0.29	0.77±0.11	0.22±0.16	0.33±0.21
mPlug-Owl2	0.74±0.25	0.26±0.20	0.35±0.25	0.61±0.31	0.27±0.24	0.35±0.29	0.59±0.28	0.29±0.31	0.34±0.32	0.78±0.23	0.19±0.20	0.27±0.26	0.82±0.13	0.17±0.11	0.27±0.16
InstructBLIP	0.70±0.23	0.32±0.29	0.39±0.32	0.78±0.10	0.25±0.24	0.32±0.30	0.46±0.26	0.31±0.24	0.36±0.26	0.75±0.24	0.18±0.19	0.26±0.25	0.71±0.16	0.15±0.16	0.20±0.18
InstructBLIP-L	0.69±0.26	0.26±0.13	0.37±0.17	0.80±0.15	0.17±0.12	0.27±0.17	0.54±0.32	0.30±0.22	0.38±0.26	0.68±0.28	0.14±0.09	0.23±0.13	0.73±0.21	0.09±0.07	0.15±0.11
LLaVA-1.5-L	0.74±0.27	0.23±0.17	0.33±0.23	0.66±0.34	0.17±0.16	0.25±0.23	0.63±0.35	0.26±0.30	0.32±0.33	0.74±0.29	0.16±0.16	0.24±0.22	0.79±0.21	0.12±0.12	0.20±0.17
LLaVA-1.5	0.76±0.25	0.23±0.20	0.32±0.27	0.65±0.33	0.17±0.16	0.25±0.23	0.61±0.32	0.22±0.28	0.28±0.30	0.75±0.26	0.19±0.19	0.27±0.26	0.81±0.13	0.14±0.11	0.22±0.16
LLaVA-1.1	0.71±0.27	0.10±0.09	0.17±0.14	0.67±0.30	0.09±0.11	0.15±0.16	0.57±0.37	0.14±0.23	0.18±0.28	0.75±0.27	0.05±0.07	0.09±0.12	0.80±0.16	0.04±0.04	0.08±0.07

Table S3. **Task-oriented performance of varied MLLMs on `PhD`.** Values after the symbol ± indicate the standard deviation across the four modes.

Question:
Is there a TV in the image?
Answer:
Yes.

Question:
Is there an armchair in the image?
Answer:
No.

Question:
Is there a fence behind the two boys sitting down in the image?
Answer:
Yes.

Question:
Is there a sports ball behind the two boys sitting down in the image?
Answer:
No.

Question:
Is the horse brown in the image?
Answer:
Yes.

Question:
Is the horse black in the image?
Answer:
No.

(a) Samples in PhD

Question:
Is there one football in this image?
Answer:
Yes.

Question:
Is there a dog in this image?
Answer:
No.

Question:
Is there direct contact between the woman and horse?
Answer:
Yes.

Question:
Is there direct contact between the woman and forest?
Answer:
No.

Question:
Is the cloud white in this image?
Answer:
Yes.

Question:
Is the cloud black in this image?
Answer:
No.

(b) Samples in AMBER

Figure S1. **Showcases of data in PhDand AMBER.** The images used in PhD are more complex and relatable to daily life, with more diverse and challenging question formats and content.

Question:
Are there exactly five cows in the image?
Answer:
Yes.

Question:
Are there exactly 3 cows in the image?
Answer:
No.

Question:
Is each of the containers rectangular in shape in the image?
Answer:
Yes.

Question:
Are the containers oval in shape in the image?
Answer:
No.

Question:
Is the man appearing sad in the image?
Answer:
Yes.

Question:
The man is usually seen with a hopeful expression while contemplating the fresh bananas that promise a delicious treat, standing beside the vibrant racks filled with ripe fruit. His presence adds to the lively atmosphere of the outdoor fruit stand, where shoppers are drawn to the tempting displays of produce. In case there is an inconsistency between the context and the image content, you should follow the image. Is the man depicted as hopeful in the image?
Answer:
No.

(a) Samples in PhD

Question:
Is there one woman in this image?
Answer:
Yes.

Question:
Are there two women in this image?
Answer:
No.

Question:
Is the sky sunny in this image?
Answer:
Yes.

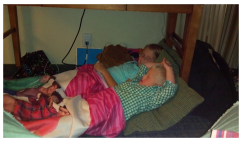
Question:
Is the sky gloomy in this image?
Answer:
No.

Question:
Is the ground muddy in this image?
Answer:
Yes.

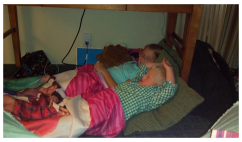
Question:
Is the ground dry in this image?
Answer:
No.

(b) Samples in AMBER

Figure S2. **Showcases of data in PhDand AMBER.** The images used in PhD are more complex and relatable to daily life, with more diverse and challenging question formats and content.



Question:
In the image, the bed frame is made of metal, providing a sturdy foundation for the children resting peacefully. The scene captures a serene moment as two kids lay down propped up on pillows, cuddling closely in the warmth of their bottom bunk. In case there is an inconsistency between the context and the image content, you should follow the image. Is the bed frame made of wood in the image?
Answer:
Yes.



Question:
Is the bed frame made of metal?
Answer:
No.

Question:
Is there a traffic light in the image?
Answer:
Yes.

Question:
Streetlamps are typically placed in close proximity to traffic signs, ensuring that the area remains well-lit for pedestrians at all times. These signs, including pedestrian crossing indications and no left turn alerts, play a crucial role in guiding traffic and enhancing safety in urban environments. In case there is an inconsistency between the context and the image content, you should follow the image. Is there a streetlamp in the image?
Answer:
No.

Question:
Is the chair made of corduroy in the image?
Answer:
Yes.

Question:
The couch is often made of leather, providing a comfortable spot for the cats to curl up together. Recently, two cats were spotted peacefully resting side by side, showcasing their bond as they enjoyed a cozy nap on the soft surface. Their contrasting fur colors added a charming touch to the serene scene. In case there is an inconsistency between the context and the image content, you should follow the image. Is the chair made of leather in the image?
Answer:
No.

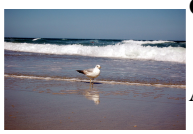
(a) Samples in PhD

Question:
Is the book open in this image?
Answer:
Yes.

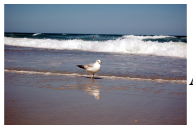
Question:
Is the book closed in this image?
Answer:
No.

Question:
Is the table clean in this image?
Answer:
Yes.

Question:
Is the table dirty in this image?
Answer:
No.



Question:
Is the sea rolling waves in this image?
Answer:
Yes.



Question:
Is the sea calm waters in this image?
Answer:
No.

(b) Samples in AMBER

Figure S3. **Showcases of data in PhDand AMBER.** The images used in PhD are more complex and relatable to daily life, with more diverse and challenging question formats and content.

Question:
Is the strawberry at the front pink?
Answer:
Yes.

Question:
Is the strawberry at the front red?
Answer:
No.

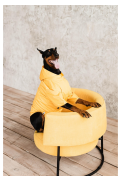
Question:
Is a red pen used to draw the blue drawing in the photo?
Answer:
Yes.

Question:
Is a red pen used to draw the red drawing in the photo?
Answer:
No.

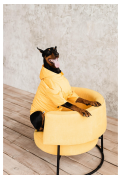
Question:
Is the woodpecker pecking at a pole?
Answer:
Yes.

Question:
Is the woodpecker pecking at a tree?
Answer:
No.

(a) Samples in PhD



Question:
Is the floor clean in this image?
Answer:
Yes.



Question:
Is the floor dirty in this image?
Answer:
No.

Question:
Is the forest lively in this image?
Answer:
Yes.

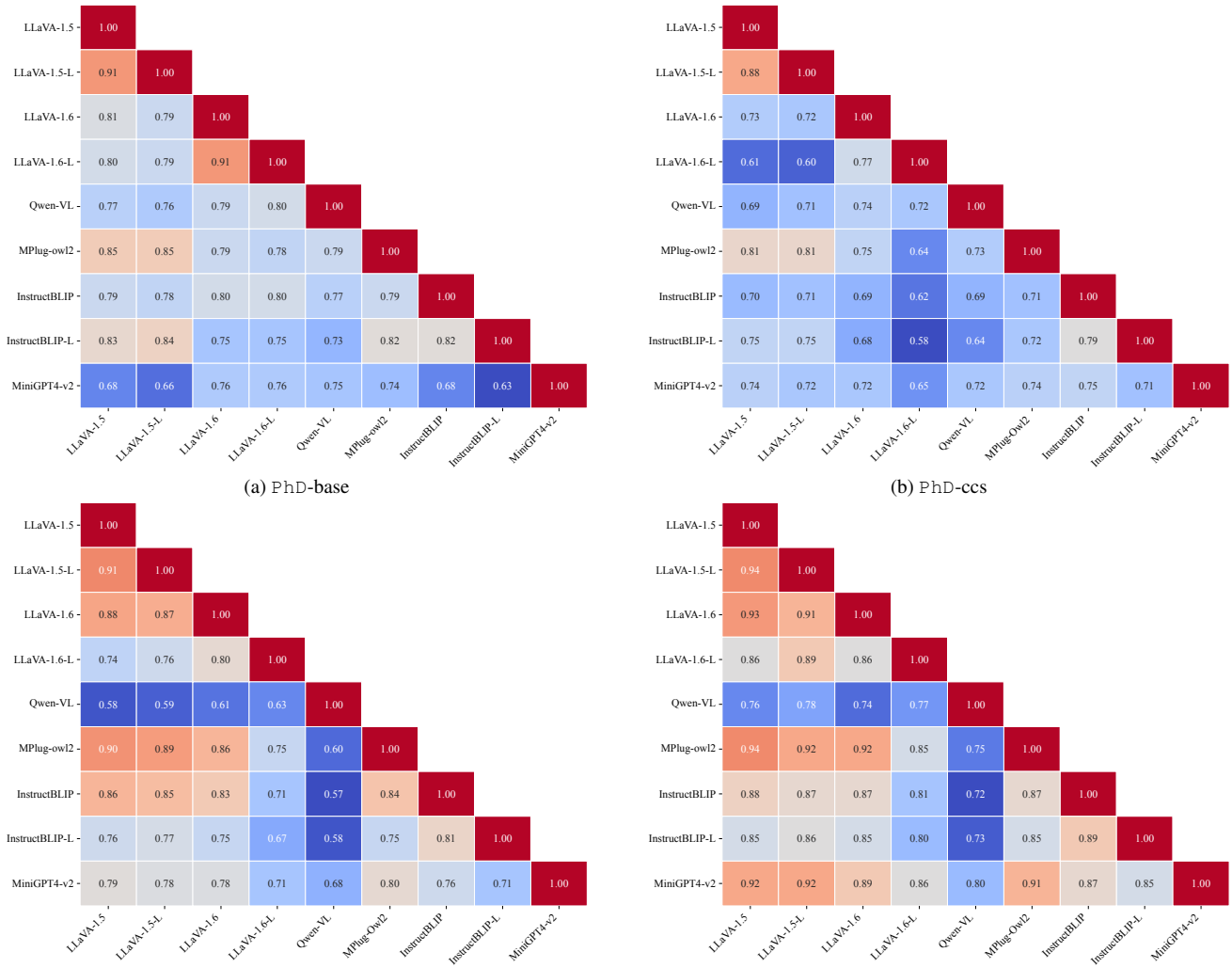
Question:
Is the forest withered in this image?
Answer:
No.

Question:
Is the cloud white in this image?
Answer:
Yes.

Question:
Is the cloud black in this image?
Answer:
No.

(b) Samples in AMBER

Figure S4. **Showcases of data in PhDand AMBER.** The images used in PhD are more complex and relatable to daily life, with more diverse and challenging question formats and content.



(a) PhD-base

(b) PhD-ccs

(c) PhD-sec

(d) PhD-icc

Figure S5. **Consistency matrices on PhD-base, PhD-ccs, PhD-sec and PhD-icc.**