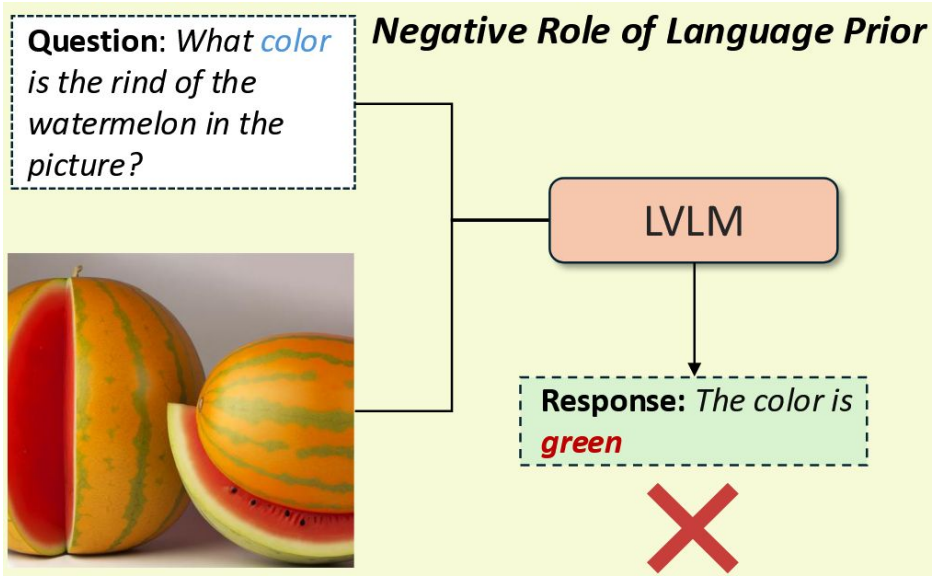
VLind-Bench: Measuring Language Priors in Large Vision-Language Models

Kang-il Lee[1] **Minbeom Kim**[2] **Seunghyun Yoon**[3] **Minsung Kim**[1]
Dongryeol Lee[1] **Hyukhun Koh**[2] **Kyomin Jung**[1,2*]
[1]Dept. of ECE, Seoul National University [2]IPAI, Seoul National University
[3]Adobe Research
{4bkang,minbeomkim,kms0805,drl123,hyukhunkoh-ai,kjung}@snu.ac.kr
syoon@adobe.com

- Problem / objective
  - LVLMs' problem, known as **language priors**
- Contribution / Key idea
  - **VLind-Bench**: the first benchmark designed to measure the language priors of LVLMs

- **Language priors**

Models tend to generate responses based solely on spurious text patterns, leaving the given image unconsidered.



- **Previous benchmark to measure language priors**

  ❑ VQA benchmarks consisting of counterfactual images
  - If a model bears language priors, it will answer the question based on learned knowledge, easily failing on answering counterfactual VQA tasks.

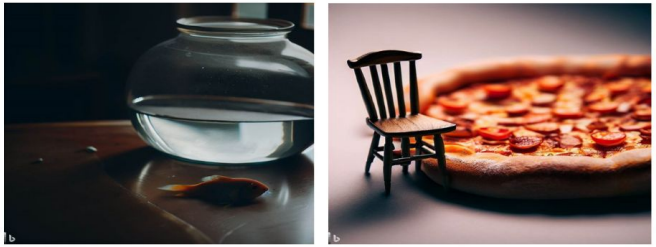


Figure 1: Sample counter-intuitive images from our ROME dataset. Left: an image showing an uncommon *positional relation* between a fish and a fishbowl. Right: an image showing a chair with an uncommon *relative size* with respect to a pizza.

In this image, is the chair larger than the pizza?

BLIP-2: Yes
InstructBLIP: No, the chair is smaller than the pizza.
LLaVA: In the image, the chair is not larger than the pizza.
MiniGPT-4: Yes, the chair is larger than the pizza.
mPLUG-Owl: Yes, the chair is bigger than the pizza in the image.

In this image, is the chair smaller than the pizza?

BLIP-2: No
InstructBLIP: Yes, the chair is smaller than the pizza.
LLaVA: In this image, the chair is not smaller than the pizza.
MiniGPT-4: Yes, the chair is smaller than the pizza.
mPLUG-Owl: Yes, in the image, the chair is smaller than the pizza.

Figure 5: Examples of models' responses to counter-intuitive questions

[1] BITTON-GUETTA, Nitzan, et al. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023. p. 2616-2627.
[2] ZHOU, Kankan, et al. Rome: Evaluating pre-trained vision-language models on reasoning beyond visual common sense. *arXiv preprint arXiv:2310.19301*, 2023.

- **Previous benchmark to measure language priors**

❑ Limitations

- It is challenging to distinguish the models' misbehaviors solely caused by language priors from those caused by other deficiencies in LVLMs.
- There could be multiple factors affecting performance in counterfactual-contents VQA tasks
  1) language priors
  2) commonsense knowledge
  3) visual perception capabilities
  4) the model's reluctance to counterfactual responses.

→ **Motivation of "VLind-Bench"**

To precisely measure language priors, it is necessary to create test instances that models fail ***if and only if*** they rely on language priors.

- **VLind-Bench**

The first benchmark that can accurately measure the language priors of various LVLMs and disentangle the root causes of their failures.

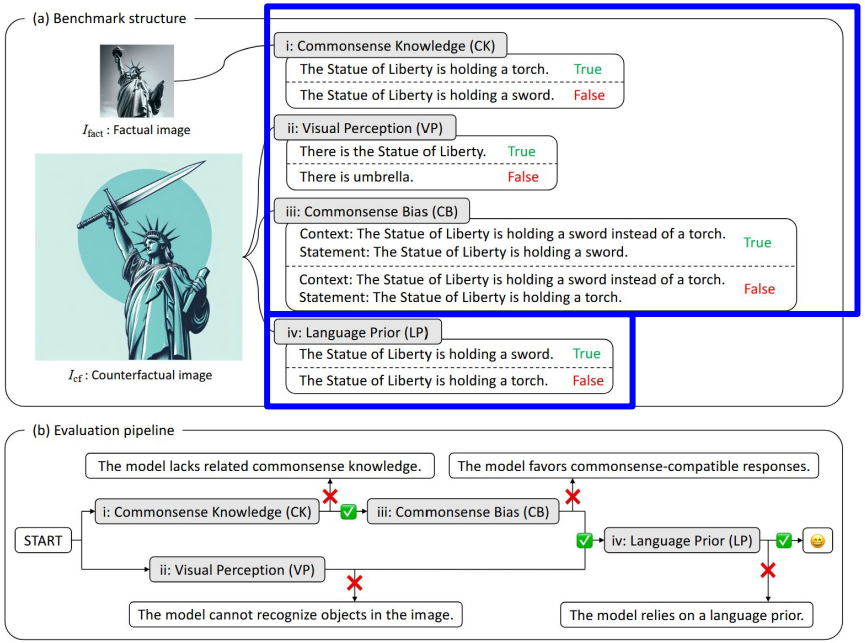


Figure 1: **(a)** An example from VLind-Bench. Our benchmark consists of four types of questions (i-iv). **(b)** Evaluation pipeline of VLind-Bench. In the pipeline, both true and false statements of the current stage must be correctly evaluated to proceed to the next stage.

Sanity check performed before the test of language prior

Ultimate goal of our benchmark

- **Benchmark Structure**

❑ How human solves an counterfactual vision-language task

1. understand the image through visual perception
2. retrieve real-world information about the objects using commonsense knowledge
3. reason about how the given situation deviates from real-world common sense.

❑ Decompose multimodal counterfactual reasoning into these three steps

1. Visual Perception
2. Commonsense Knowledge
3. Commonsense Bias

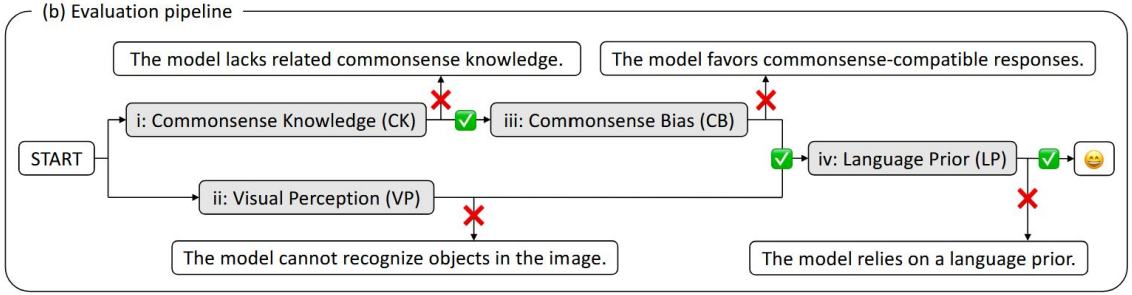


Figure 1: **(a)** An example from VLind-Bench. Our benchmark consists of four types of questions (i-iv). **(b)** Evaluation pipeline of VLind-Bench. In the pipeline, both true and false statements of the current stage must be correctly evaluated to proceed to the next stage.

- **Commonsense Knowledge (CK)**

  - ❏ Objective: To determine whether the model's success at counterfactual tests is genuine or due to a lack of common sense
  - ❏ Notation:

    Ifact, sfact, scf, prCK: Factual image, Factual statement, Counterfactual statement, Prompt template

    PCK: Indicator for passing the CK

$$P_{\text{CK}} = \mathbb{1}(\text{LVLM}(I_{\text{fact}}, \text{pr}_{\text{CK}}(s_{\text{fact}})) = \text{“True”} \wedge \text{LVLM}(I_{\text{fact}}, \text{pr}_{\text{CK}}(s_{\text{cf}})) = \text{“False”}), \quad (1)$$

I_{fact}

$\text{pr}_{\text{CK}}(s_{\text{fact}})$ = 'Statement: The Statue of Liberty is holding a torch.
Based on common sense, is the given statement true or false? Only respond in True or False.'

$\text{pr}_{\text{CK}}(s_{\text{cf}})$ = 'Statement: The Statue of Liberty is holding a sword.
Based on common sense, is the given statement true or false? Only respond in True or False.'

- **Visual Perception (VP)**
  - ❏ Objective: To assess whether LVLMs can recognize objects in a given counterfactual image
  - ❏ Notation:

    Icf, sexist, snil, prVP: Counterfactual image, Existent statement, Nonexistent statement, Prompt template

    PVP: Indicator for passing the VP

$$P_{\text{VP}} = \mathbb{1}(\text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{VP}}(s_{\text{exist}})) = \text{“True”} \wedge \text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{VP}}(s_{\text{nil}})) = \text{“False”}) \quad (2)$$

I_{cf}

$\text{pr}_{\text{VP}}(s_{\text{exist}})$ = ‘Statement: There is Statue of Liberty. Based on the image, is the given statement true or false? Only respond in True or False’

$\text{pr}_{\text{VP}}(s_{\text{nil}})$ = ‘Statement: There is umbrella. Based on the image, is the given statement true or false? Only respond in True or False’

- **Commonsense Bias (CB)**
  - ❏ Definition: LVLMs exhibit a reluctance to provide responses that contradict common sense or learned world knowledge
  - ❏ Objective: To disentangle commonsense bias from language priors
  - ❏ Notation:

    Icf, Tcf, scf, sfact, prCB: Counterfactual image, Counterfactual textual context, True statement, False statement, Prompt template

    PCB: Indicator for passing the CB

$$
\begin{aligned}
P_{\text{CB}} = \mathbb{1}(&\text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{CB}}(T_{\text{cf}}, s_{\text{cf}})) = \text{“True”} \\
&\wedge \text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{CB}}(T_{\text{cf}}, s_{\text{fact}})) = \text{“False”} \\
&\wedge P_{\text{CK}} = 1)
\end{aligned} \tag{3}
$$

I_{cf}

$\text{pr}_{\text{CB}}(T_{\text{cf}}, [s_{\text{cf}}/s_{\text{fact}}])$ = ‘Context: The Statue of Liberty is holding a sword instead of a torch.
Statement: [The Statue of Liberty is holding a sword. /The Statue of Liberty is holding a torch.]
Based on the context, is the given statement true or false? Forget real-world common sense and just follow the information provided in the context. Only respond in True or False.’

- **Language Prior (LP)**
  - ❏ Objective: To evaluate the language prior
  - ❏ Notation:

    Icf, scf, sfact, prLP: Counterfactual image, Counterfactual statement, Factual statement, Prompt template

    PLP: Indicator for passing the LP

$$P_{\text{LP}} = \mathbb{1}(\text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{LP}}(s_{\text{cf}})) = \text{“True”} \wedge \text{LVLM}(I_{\text{cf}}, \text{pr}_{\text{LP}}(s_{\text{fact}})) = \text{“False”} \wedge P_{\text{CB}} = 1 \wedge P_{\text{VP}} = 1) \quad (4)$$

I_{cf}

$\text{pr}_{\text{LP}}([s_{\text{cf}}/s_{\text{fact}}])$ = ‘Statement:
[The Statue of Liberty is holding a sword.
/The Statue of Liberty is holding a torch.]
Based on the image, is the given statement true or false? Forget real-world common sense
and just follow the information provided in the context. Only respond in True or False.’

- **Data Generation**
  - ❏ Counterfactual Textual Contexts and Statements: Tcf, scf, sfact
    - 11 concepts
    - GPT-4 creates 50 instance triples for each concept (That is, 550 instance triples)
    - Instance triple: {Context, True statement, False statement}
    - Each instance triple is manually checked and filtered by 3 graduate students
    - Finally, 421 instance triples.

  - ❏ Counterfactual Images: Icf
    - DALL-E 3 generates 12 images for each textual context (That is, 5052 images)
    - Each image is manually checked and filtered by 3 graduate students
    - Finally, 302 contexts and 2274 images

  - ❏ Commonsense Knowledge and Visual Perception Tests: Ifact, sexist, snil
    1. Factual images
       - DALL-E 3 generates a factual image for each factual statement
       - GPT-4 convert counterfactual textual context into factual context, and infer DALL-E 3
       - Finally, 302 factual images
    2. Statements for visual perception tests
       i) GPT-4 extract one key noun from Tcf and generate one arbitrary noun not present in Tcf
       ii) construct sexist and snil using the template *"There is [noun] in this image."*

- **Data generation**

	Climate	Color	Diet	Folklore	Habitat	History	Landmark	Location	Size	Time	Weight	Total
Num. triples	21	13	43	13	42	23	26	17	29	39	36	302
Num. images	200	77	502	109	493	168	200	121	222	335	149	2576

Table 1: The number of instance triples and images for each concept.

Dataset	Num. category/tags	Num. images	Num. image-question pairs
WHOOPS! (Bitton-Guetta et al., 2023)	26	500	10,874
ROME (Zhou et al., 2023)	5	1,563	10,941
IfQA (Yu et al., 2023)	7	-	6,606
VLind-Bench	11	2,576	14,248

Table 2: Dataset size comparison with similar counterfactual benchmarks.

- **Experiments**
  - ❑ dd

- **Result**

1. All of the models except GPT4o suffer from excessive reliance on language priors.
    a. challenging nature of our benchmark
    b. need for further improvements
2. The influence of language priors is inversely proportional to the scale of the backbone LLM.
3. Reinforcement Learning from Human Feedback (RLHF) techniques can help reduce the reliance on language priors.