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

Abstract

Large Vision-Language Models (LVLMs) have demonstrated outstanding performance across various multimodal tasks. However, they suffer from a problem known as *language prior*, where responses are generated based solely on textual patterns while disregarding image information. Addressing the issue of language prior is crucial, as it can lead to undesirable biases or hallucinations when dealing with images that are out of training distribution. Despite its importance, current methods for accurately measuring language priors in LVLMs are poorly studied. Although existing benchmarks based on counterfactual or out-of-distribution images can partially be used to measure language priors, they fail to disentangle language priors from other confounding factors. To this end, we propose a new benchmark called VLind-Bench, which is the first benchmark specifically designed to measure the language priors, or *blindness*, of LVLMs. It not only includes tests on counterfactual images to assess language priors but also involves a series of tests to evaluate more basic capabilities such as commonsense knowledge, visual perception, and commonsense biases. For each instance in our benchmark, we ensure that all these basic tests are passed before evaluating the language priors, thereby minimizing the influence of other factors on the assessment. The evaluation and analysis of recent LVLMs in our benchmark reveal that almost all models exhibit a significant reliance on language priors, presenting a strong challenge in the field.[1]

1 Introduction

Recent Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across various tasks through pre-training on massive multimodal datasets and visual instruction tuning. (Liu et al., 2023; Dai et al., 2023a; Zhu et al., 2024; Ye et al., 2024; Peng et al., 2023). However, these models tend to generate responses based solely on spurious text patterns, leaving the given image unconsidered. We refer to this problem as *language prior*, borrowing the term from the Visual Question Answering (VQA) community (Agrawal et al., 2018). Such language priors can lead to undesirable biases (Hall et al., 2023) and hallucinations (Wang et al., 2023). For example, when a model is presented with an image of a red banana and a yellow apple along with the question, "Is the banana yellow?," it has been observed that the model frequently responds with "Yes," ignoring the image content (Zhou et al., 2023). To develop a trustworthy LVLM, resolving the language prior issue is crucial; however, it has not been explored much nor has benchmarks that can accurately measure the issues.

One approach to measure language priors is assessing performance on VQA benchmarks consisting of counterfactual images (e.g., WHOOPS! (Bitton-Guetta et al., 2023) and ROME (Zhou et al., 2023)). If a model bears language priors, it will answer the question based on learned facts or common sense from its parametric knowledge without collaborating information in the given context (i.e., image); easily failing on answering counterfactual VQA tasks. However, it is challenging to distinguish the models' misbehaviors solely caused by language priors from those caused by other deficiencies in LVLMs. For example, there could be multiple factors affecting performance in counterfactual-contents VQA tasks – not only language priors but also commonsense knowledge, visual perception capabilities, and the model's reluctance to counterfactual responses. Such confounding factors make it difficult to evaluate methodologies for improving language prior problems and to assess progress in the research field.

In this paper, we propose VLind-Bench, the first

*Corresponding authors.

[1]Evaluation code and benchmark data are available at https://github.com/klee972/vlind-bench.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 4129–4144

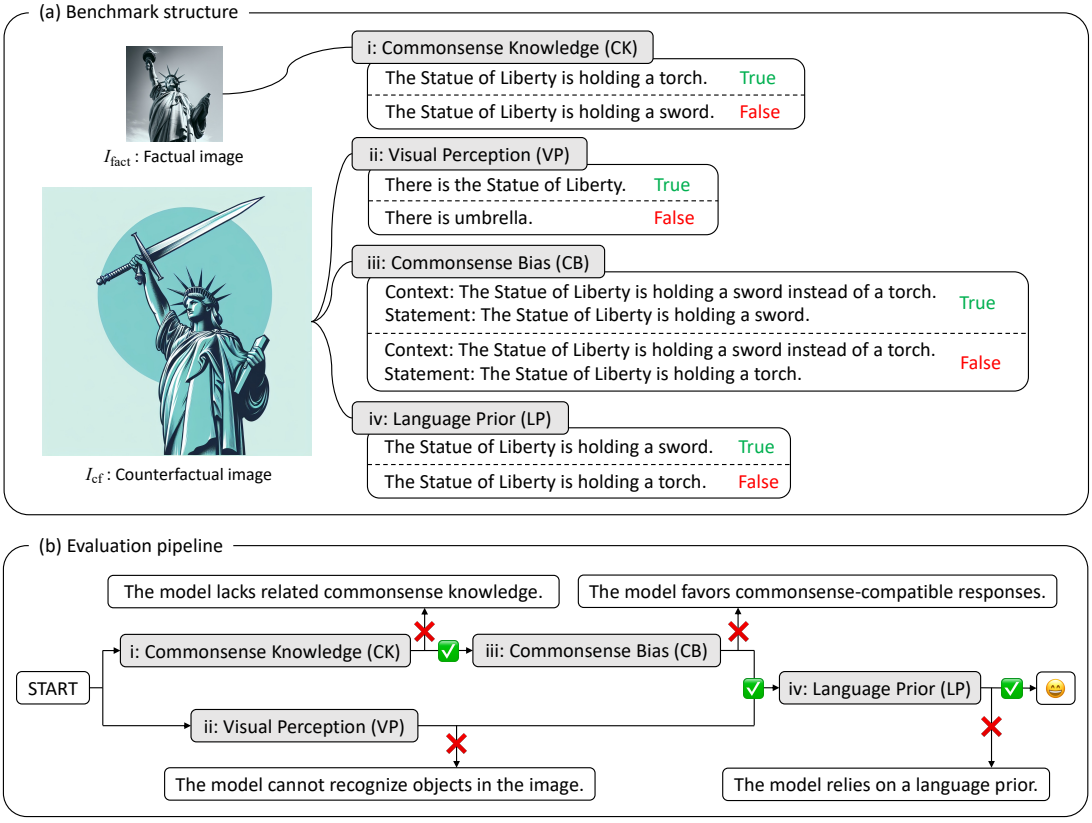


Figure 1: **(a)** An example from VLind-Bench. Our benchmark consists of four types of questions (i-iv). **(b)** Evaluation pipeline of VLind-Bench. In the pipeline, both true and false statements of the current stage must be correctly evaluated to proceed to the next stage.

benchmark that can accurately measure the language priors, or *blindness*, of various LVLMs and disentangle the root causes of their failures. To precisely measure language priors, it is necessary to create test instances that models fail **if and only if** they rely on language priors. For this purpose, we meticulously design a sequence of tests and measure the accuracy on each of them (Figure 1 (a)). Specifically, each instance in VLind-Bench involves four tests that can check whether a model possesses (1) commonsense knowledge, (2) visual perception, (3) commonsense bias, and (4) language prior. The first three serve as a sanity check performed before the test of language prior, which is the ultimate goal of our benchmark (Figure 1 (b)). To the best of our knowledge, existing benchmarks can only show the individual task-level performance of LVLMs.

With VLind-Bench, we evaluate recent open-source and proprietary LVLMs' language priors. The results show that all of the models except GPT-4o (OpenAI, 2024) suffer from excessive reliance on language priors, demonstrating the challenging nature of our benchmark and the need for further improvements. Furthermore, our experiment and analysis on existing LVLMs show that the influence of language priors is inversely proportional to the scale of the backbone LLM. We also reveal that Reinforcement Learning from Human Feedback (RLHF) techniques (Yu et al., 2024a,b), which are designed to mitigate hallucinations, can help reduce the reliance on language priors.

2 Related Work

2.1 Large Vision-Language Models

Recently, there has been a lot of effort in extending Large Language Models (LLMs) to include visual inputs, forming a new class of models known as Large Vision-Language Models (LVLMs) (Liu et al., 2023; Dai et al., 2023a; Zhu et al., 2024; Ye et al., 2024; OpenAI, 2023b, 2024; Google, 2024). These LVLMs are gaining attention as a new paradigm in vision-language learning by transferring the exceptional properties of LLMs, such as multi-step reasoning ability and in-context learning, to the multimodal domain. However, these LVLMs are not free from the bias and hallucination issues inherent in LLMs (Hall et al., 2023; Li

et al., 2023; Gunjal et al., 2024; Zhou et al., 2024a; Dai et al., 2023b; Min et al., 2024). Despite this, creating benchmarks to diagnose these problems is more challenging with the image modality, leading to slower progress in benchmark development compared to LLMs.

2.2 Benchmarks with Counterfactual Context

Since counterfactual contexts can assess the robustness and generalization capabilities of LLMs or LVLMs, several benchmarks utilizing this approach have been proposed. These benchmarks assume that if a model responds based on memorized facts without properly understanding the context of text or images, it would fail to correctly solve tasks conditioned on counterfactual contexts. Benchmarks such as IfQA (Yu et al., 2023) and DisentQA (Neeman et al., 2023) counterfactually augment textual contexts to determine whether the language model accurately incorporates augmented information when answering questions. Wu et al. (2024) evaluate LLMs on reasoning tasks based on counterfactual contexts. Benchmarks like WHOOPS! (Bitton-Guetta et al., 2023), ROME (Zhou et al., 2023), HALLUSIONBENCH (Guan et al., 2024), and ViLP (Luo et al., 2024) evaluate the counterfactual reasoning abilities of multimodal models by conducting VQA tasks conditioned on counterfactual images. However, these benchmarks cannot disentangle the reliance on language priors and commonsense biases of a model.

3 Benchmark Structure

VLind-Bench conducts four types of assessments, each designed to test different capabilities, as illustrated in Figure 1 (a). By providing multiple tests concerning the **exact** same image or text that are used in the language prior test, it is possible to check if the model has the essential abilities to make the language prior test meaningful. Depending on the problem's characteristics, each test utilizes one of two images, either factual or counterfactual, as input.

First, we provide a counterfactual image along with two statements and evaluate whether the model can correctly classify these statements as true or false based on the image (Figure 1 (a) - iv: Language Prior). If the model relies on language priors, it will not incorporate the counterfactual circumstances presented in the image into its reasoning, achieving low performance on this test.

However, merely answering questions about counterfactual images is insufficient to accurately measure the language priors due to several confounding factors. Firstly, when a model fails a task involving a counterfactual image, it is unclear whether this failure is due to the model's reliance on language priors or because the model possesses *commonsense bias*. Here, commonsense bias refers to the tendency of models, including unimodal language models, to avoid responding in ways that contradict common sense. Therefore, we evaluate whether the model can overcome such commonsense bias *regardless of modality*, by providing the model with the image and *a text description of the image* as input (Figure 1 (a) - iii: Commonsense Bias).

Additionally, the failure in the counterfactual task might stem from an inability to recognize the objects in the counterfactual image. Conversely, the model may simply lack common sense and pass the test merely by chance. To this end, we provide two tests to check commonsense knowledge and visual perception abilities. The statements used for checking commonsense knowledge are identical to those for language priors, but *factual* images are given instead of counterfactual images, and the models are instructed to evaluate the truth values based on common sense (Figure 1 (a) - i: Commonsense Knowledge). In the case of visual perception, counterfactual images are still used; however, the statements are designed to assess the model's ability to recognize objects (Figure 1 (a) - ii: Visual Perception).

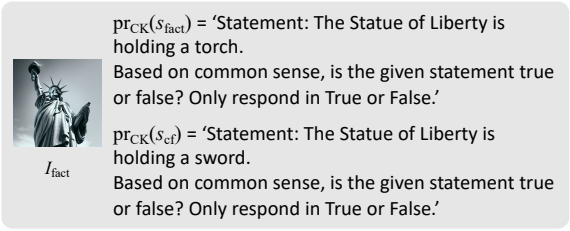
If we introspect how a human solves an counterfactual vision-language task, the three additional assessment types we propose appear more convincing. Humans first understand the image through visual perception, then retrieve real-world information about the objects using commonsense knowledge, and finally reason about how the given situation deviates from real-world common sense. Decomposing multimodal counterfactual reasoning into these three steps is a very natural approach, and each of these steps directly corresponds to our Visual Perception, Commonsense Knowledge, and Commonsense Bias tasks.

If a model fails any test assessing its basic ability, evaluating it on more complex tests that rely on that basic ability would be meaningless. Therefore, the evaluation of our benchmark proceeds sequentially, starting with easier problems that assess fundamental abilities and gradually advancing to

more difficult problems that are counterfactual and multimodal in nature (Figure 1 (b)). This pipelined evaluation paradigm could be more universally applied, not only for measuring language priors but also for more accurately assessing the varying capabilities of AI systems.

3.1 Commonsense Knowledge (CK)

First, it is essential to verify whether the model possesses commonsense knowledge about the instances of the benchmark. This step allows us to determine whether the model's success at counterfactual tests is genuine or due to a lack of common sense. Therefore, we introduce a Commonsense Knowledge test (CK) to assess the model's commonsense knowledge about the given instances. Specifically, the CK comprises one image I_{fact} and two statements s_{fact} and s_{cf}. The image I_{fact} depicts a factual circumstance that aligns with common sense (e.g., an image of the Statue of Liberty). Among the two statements, s_{fact} is a factual statement that is true based on real-world common sense (e.g., "The Statue of Liberty is holding a torch."), while s_{cf} is a counterfactual statement that is false (e.g., "The Statue of Liberty is holding a sword."). Also, we use the prompt template, pr_{CK}, to instruct the LVLM to evaluate the truth value of the input text based on common sense.



To pass the CK, the model must accurately predict the truth value of both statements:

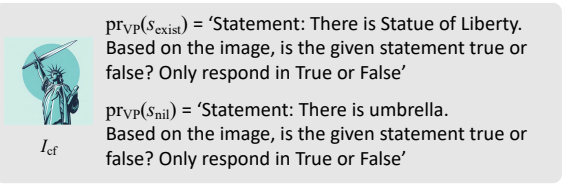
$$P_{\text{CK}} = \mathbb{1}(\text{LVLM}(I_{\text{fact}}, \text{pr}_{\text{CK}}(s_{\text{fact}})) = \text{“True”} \wedge \text{LVLM}(I_{\text{fact}}, \text{pr}_{\text{CK}}(s_{\text{cf}})) = \text{“False”}), \quad (1)$$

where P_{CK} indicates whether the model passed CK or not. $\text{LVLM}(i, t)$ is a composition of two functions: one that maps the image input i and text input t to the LVLM's response, and another that maps the LVLM's response to "True" or "False" using a string match.

3.2 Visual Perception (VP)

The fundamental ability underpinning all multimodal tasks is visual perception, particularly the ability to recognize objects (Locatello et al., 2020; Burgess et al., 2019). Similar to the CK, evaluating a model on more complex tasks would be meaningless when it fails in object recognition. Therefore, we introduce the Visual Perception test (VP) to assess whether LVLMs can recognize objects in a given counterfactual image. VP consists of one counterfactual image I_{cf} and two statements s_{exist} and s_{nil}. Contrary to the CK, the image I_{cf} shows a counterfactual scene, which contradicts the world knowledge or common sense (e.g., an image of the Statue of Liberty holding a sword). The reason for using counterfactual images is that the VP needs to evaluate visual perception capabilities on the same images that are used for language prior assessments, where the use of counterfactual images is essential.

In VP, both the two statements say that "There is *object* in the image.", while the objects are set such that s_{exist} is true and s_{nil} is false under the given image (e.g., "There is the Statue of Liberty." and "There is umbrella."). To this end, we define P_{VP} to indicate whether the model passed VP, with a prompt template pr_{VP} to instruct the models to evaluate the truth value of input text based on the given image.



The indicator for passing the VP, P_{VP}, is defined similarly:

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

3.3 Commonsense Bias (CB)

It has been observed that LVLMs, including LLMs, exhibit a reluctance to provide responses that contradict common sense or learned world knowledge, even when they are explicitly instructed to respond based on counterfactual contexts (Bitton-Guetta et al., 2023; Zhou et al., 2023; Neeman et al., 2023; Yu et al., 2023). We propose a Commonsense Bias test (CB) to disentangle this bias from language priors, which is the goal of this benchmark. To eliminate the influence of modality in the evaluation of commonsense bias, we provide LVLMs

with a counterfactual textual context T_{cf} and a counterfactual image I_{cf} as input. Also, we provide the models with two statements, s_{cf} and s_{fact}, which are true and false respectively under the given context. We wrap the context and statement with a prompt template pr_{CB}, which instructs the model to explicitly follow the information provided in the context, rather than common sense.

$\text{pr}_{\text{CB}}(T_{\text{cf}}, [s_{\text{cf}}/s_{\text{fact}}])$ = '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.'

I_{cf}

The indicator for CB is as follows:

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

Note that $P_{\text{CB}} = 1$ only if $P_{\text{CK}} = 1$, according to the proposed evaluation pipeline (Figure 1 (b)).

3.4 Language Prior (LP)

The evaluation of the language prior, which is the final and most crucial issue, is conducted through the Language Prior test (LP) involving a counterfactual image I_{cf} and two statements s_{cf} and s_{fact}. Basically, the LP is nearly identical to the CB in all aspects except for the absence of text context T_{cf} and a slight difference in prompt template pr_{LP}.

$\text{pr}_{\text{LP}}([s_{\text{cf}}/s_{\text{fact}}])$ = '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.'

I_{cf}

The indicator for LP is as follows:

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

4 Data Generation

Here, we explain the data generation process of VLind-Bench. As described in the previous section, the benchmark consists of four types of tests, incorporating various forms of images and texts. First, at the core of the benchmark data, there are counterfactual textual context T_{cf} and image I_{cf}, accompanied by two statements s_{cf} and s_{fact}, for CB and LP. To evaluate CK and VP, there are also a factual image I_{fact} and two statements s_{exist} and s_{nil} regarding object recognition. To ensure the high quality of the data samples, we proceed with the following procedure.

Counterfactual Textual Contexts and Statements First, we generate counterfactual textual context T_{cf} and corresponding statements s_{cf} and s_{fact}, which are true and false, respectively, based on the context. The contexts must describe a wide range of real-world topics and be suitable for visual depiction. To achieve this goal, we selected 11 concepts that span various aspects of commonsense knowledge, ranging from natural sciences such as `climate` and `habitat`, to humanities such as `history` and `landmark`.

For each selected concept, we employed GPT-4 (`gpt-4-0125-preview`) (OpenAI, 2023a) to create 50 *instance triples*, each consisting of a context, a true statement, and a false statement. We provided a detailed instruction with 3-shot prompt as input, using hand-crafted concept-specific examples to reflect the characteristics of each concept.

To ensure the quality of the generated data, three graduate students manually checked the correctness of the triples. We then conducted a majority vote among the three annotations to determine whether each triple should remain in our benchmark. As a result, the initial set of 550 instance triples was reduced to 421.

Counterfactual Images Next, we proceed with the generation of counterfactual image I_{cf} from the filtered textual contexts. Given the significance of LP in our benchmark, we generate multiple images per test for LP, unlike factual images where we generate only one image per test. The images are generated using DALL-E 3 (OpenAI, 2023c), where the textual context T_{cf} is provided as input, and 12 images are sampled. To provide diversity of image style, we produce four images each in photorealistic, illustration, and cartoon styles per one textual context. Consequently, for the 421 contexts, a total of 5,052 images are generated.

The generated images must provide sufficient context to accurately classify the statements as true or false and be free of artifacts. Similar to the previous stage, each image is verified by three graduate

	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.

student reviewers and filtered using a majority vote. Contexts with no accepted images are also filtered at this stage. After this filtering process, 302 contexts and 2,274 images remained in the benchmark dataset.

Commonsense Knowledge and Visual Perception Tests In the final stage of data generation, we produce factual images I_{fact} for CKs and statements s_{exist} and s_{nil} for VPs. For the factual image, since it needs to describe a circumstance where s_{fact} as true, we input s_{fact} directly into DALL-E 3 to generate the image. However, some s_{fact}'s are very difficult to translate into images using this method. In such cases, we convert T_{cf} into factual textual context using GPT-4, or alternatively, we use existing images from the web.

Statements for visual perception tests are simply sentences about the presence of objects and thus can be generated using a template. We first prompt GPT-4 to extract one key noun from T_{cf} and generate one arbitrary noun not present in T_{cf}. Then, we construct s_{exist} and s_{nil} using the template "There is [*noun*] in this image.".

To verify the quality of the generated I_{fact}, s_{exist}, and s_{nil}, we evaluate whether OpenAI GPT-4o (`gpt-4o-2024-05-13`) (OpenAI, 2024), which is the most advanced available LVLM, passes the CK and VP. For instances where GPT-4o fails, human verification was conducted. If the failure was due to an error in the data generation process, we addressed the cause of the error by either regenerating the factual image or manually correcting the nouns in statements.

Details for human verification and input prompts are provided in Appendix A.

Statistics The statistics of the benchmark data generated through the process are presented in Table 1. The difficulty of data generation varies for each concept, resulting in different proportions of samples being filtered out during the human review process. Ultimately, a total of 302 instance triples and 2,576 images, encompassing both counterfactual and factual images, were included in the benchmark. We compare the size of VLind-Bench with other counterfactual benchmarks in Table 2. Data samples for each concept can be found in Appendix E.

5 Experiments

5.1 Metrics

In section 3, all indicator values for the four tests have been defined for a single instance. For some test $\mathcal{T} \in \{\text{CK}, \text{VP}, \text{CB}, \text{LP}\}$, the final VLind-Bench score $S_{\mathcal{T}}$, is represented as the average of the indicator values $P^i_{\mathcal{T}}$'s across all instances that have passed previous tests.

$$S_{\mathcal{T}} = \frac{1}{M_{\mathcal{T}}} \sum_{i=1}^{N} P^i_{\mathcal{T}} \tag{5}$$

Here, i is the data index, N is the number of total instances in our benchmark, and $M_{\mathcal{T}}$ is the number of instances that have passed all the previous tests before $\mathcal{T}$ (which is essentially the number of instances considered by $\mathcal{T}$). To be more concise, $M_{\text{CK}} = M_{\text{VP}} = N$, $M_{\text{CB}} = |\{i \mid P^i_{\text{CK}} = 1\}|$ and $M_{\text{LP}} = |\{i \mid P^i_{\text{CB}} = 1 \wedge P^i_{\text{VP}} = 1\}|$. We refer to these four scores as *pipeline scores*, as they reflect the pipelined evaluation structure of VLind-Bench (columns under "Pipeline Score" in Table 3). Alternatively, following the common definition of accuracy, the performance can be expressed as the ratio of correct instances to the total number of instances (columns under "Accuracy" in Table 3).

Models	Pipeline Score				Accuracy	
	S_{CK}	S_{VP}	S_{CB}	S_{LP}	CB	LP
Proprietary LVLMs						
GPT-4o (OpenAI, 2024)	**93.0**	**96.0**	**96.8**	**89.8**	**97.0**	**89.4**
GPT-4V (OpenAI, 2023b)	90.1	85.4	90.8	77.6	91.1	75.6
Gemini Pro Vision (Google, 2024)	80.5	90.4	77.0	79.0	75.5	65.5
Open-source LVLMs						
LLaVA-NEXT 72B (Qwen 1.5 72B Chat) (Li et al., 2024)	**94.4**	95.7	76.1	58.6	75.5	46.7
LLaVA-NEXT 34B (Nous Hermes 2 Yi 34B) (Li et al., 2024)	80.5	85.8	61.7	61.1	67.2	44.5
LLaVA-1.5 13B (Vicuna v1.5 13B) (Liu et al., 2024)	59.9	92.1	40.9	42.0	31.5	20.9
LLaVA-1.5 7B (Vicuna v1.5 7B) (Liu et al., 2024)	0.0	0.0	-	-	0.0	0.0
+ RLAIF-V (Yu et al., 2024b)	17.9	8.3	48.1	25.0	54.3	35.7
InstructBLIP 13B (Dai et al., 2023a)	66.6	79.5	54.2	57.8	46.7	28.0
InstructBLIP 7B (Dai et al., 2023a)	58.6	73.5	28.2	14.6	27.2	21.0
OmniLMM 12B (Zephyr 7B β) (Yu et al., 2024a)	88.1	97.7	**78.6**	**81.4**	**79.5**	**66.4**
MiniCPM-V-2 2.8B (Yu et al., 2024a)	76.2	**98.3**	56.5	68.1	49.0	34.1
Backbone LLMs						
Qwen 1.5 72B Chat (Bai et al., 2023)	75.8	-	69.9	-	74.2	-
Nous Hermes 2 Yi 34B (NousResearch, 2023)	**83.1**	-	75.3	-	**77.8**	-
Vicuna v1.5 13B (Team, 2023)	57.9	-	**80.0**	-	69.2	-
Vicuna v1.5 7B (Team, 2023)	0.0	-	-	-	0.0	-
Zephyr 7B β (Tunstall et al., 2023)	62.3	-	45.7	-	40.7	-

Table 3: Main experimental results on VLind-Bench.

5.2 Models

We have selected and evaluated recent proprietary and open-source LVLMs on the VLind-Bench. The open-source LVLMs were chosen to represent a diverse range of scales and training methodologies. Unfortunately, the performance of the InstructBLIP models could not be evaluated using the prompt template from section 3, as they completely failed to generate responses. Therefore, we utilized a modified prompt, in which the question sentence was placed at the end. Additionally, we assessed the performance of some backbone LLMs on CK and CB tasks without the image input. To ensure the reproducibility of the experiments, all inferences were conducted under a zero temperature setting. All the experiments are conducted using 4 NVIDIA RTX A6000 GPUs.

5.3 Main Results

The overall model performance is shown in Table 3. Surprisingly, numerous models demonstrated somewhat low scores in S_{CK}, implying a deficiency of commonsense knowledge in LVLMs. Conversely, S_{VP} scores concerning object recognition ability exhibited relatively high scores. This pattern of low commonsense knowledge scores and high visual perception scores aligns with observations from previous work (Zhou et al., 2023). Additionally, the lower S_{CB} and CB scores compared to S_{CK} indicate that LVLMs are reluctant to respond contrary to commonsense knowledge.

When comparing LP and S_{LP} scores, it is evident that some models with similar LP scores exhibit differing S_{LP} scores. For instance, while the LLaVA 1.5 13B model and the InstructBLIP 7B model have similar LP scores, the LLaVA model achieves nearly three times higher S_{LP} score. This clear lack of correlation between LP and S_{LP} scores indicates that our pipelined evaluation provides additional information beyond what can be obtained by conducting task-level evaluation alone.

Finally, the generally low S_{LP} score suggests that all models, except for GPT-4o, exhibit a reliance on language priors. This reliance was more pronounced in open-source models compared to proprietary ones. The reliance on language priors appeared inversely proportional to the scale of the backbone LLM. This trend can be observed by comparing the S_{LP} scores across various sizes of models within the same LLaVA and InstructBLIP series.

To verify the validity of the VLind-Bench, we conducted experiments on a small handcrafted evaluation set, and the results are provided in Appendix D.

RLHF-V An exception to such trend between model scale and language prior is the superior performance of models that applied the RLHF-V (Yu et al., 2024a) methodologies. Models such as OmniLMM and MiniCPM trained using RLHF-V,

Model (Score Type)	Climate	Color	Diet	Folklore	Habitat	History	Landmark	Location	Size	Time	Weight	Total
GPT-4o (S_{CK})	95.2	76.9	97.7	61.5	92.9	100.0	84.6	88.2	93.1	100.0	100.0	93.0
GPT-4o (S_{LP})	83.3	93.3	97.1	91.2	98.2	92.0	69.7	100.0	99.2	100.0	61.0	89.8
OmniLMM (S_{CK})	100.0	84.6	97.7	76.9	92.9	87.0	92.3	82.4	41.4	100.0	94.4	88.1
OmniLMM (S_{LP})	73.7	81.9	99.0	87.8	86.7	88.2	47.9	98.2	45.5	80.7	0.0	81.4

Table 4: Performance of selected models for different concepts.

demonstrated superior performance compared to models of similar or greater scale. Specifically, RLHF-V employs a method called Dense Direct Preference Optimization (DDPO) to mitigate multimodal hallucination. DDPO constructs win-lose pairs by having humans modify only the hallucinatory spans in the model responses to align with image information, thereby forcing the use of visual modality to increase the reward. Such construction of training data might be the reason for the reduced reliance on language prior. Additionally, the high performance of these methods on counterfactual images suggests that the ability to utilize image information generalizes to out-of-distribution samples. Applying RLAIF-V (Yu et al., 2024b), an AI-feedback variant of RLHF-V, to LLaVA 1.5 7B also results in significant performance improvement.

LLM performance Some might question whether the performance of LVLM is significantly influenced by the performance of its backbone LLM. To answer this question, we conducted an evaluation of several backbone LLMs on CK and CB tasks. The results, as illustrated in columns S_{CK} and S_{CB}, indicate that the performance of the LLMs is not highly correlated to the performance of the LVLMs. Consequently, we can conclude that the absolute scale of the backbone LLMs and the training methodology have a more substantial impact on the final performance of LVLMs than the performance of the backbone LLMs themselves.

Another finding is that the LVLMs are sometimes superior to their original backbone LLMs on S_{CB}. Given that S_{CB} encompasses the same content in both image and text formats, this suggests that, in certain scenarios, learning from the visual modality may enhance robustness in the text modality.

6 Discussion

Performance by Concept One particularly interesting finding is that the model performance varies significantly depending on the concept. For instance, high-performing open-source models such as OmniLMM scored zero in S_{LP} for the concept of "weight," and even GPT-4o only managed to achieve a score of 61.0% (Table 4). This suggests that although LVLMs might possess real-world knowledge about physical properties like weight, they lack robust concepts of these properties that can be generalized under counterfactual situations.

Chain-of-Thought Prompting LLMs are known to respond more comprehensively by generating intermediate reasoning steps. In this section, we assess the effect of Zero-shot-CoT (Kojima et al., 2022) on VLind-Bench tasks by replacing the instruction in our prompts "Only respond in True or False." to "Let's think step by step.".

	Pipeline Score				Accuracy	
Models	S_{CK}	S_{VP}	S_{CB}	S_{LP}	CB	LP
Zero-shot-CoT						
GPT-4o	91.4	94.7	93.1	89.4	93.0	87.8
GPT-4V	91.4	95.7	93.1	87.1	92.7	85.0
LLaVA-NEXT 72B	94.0	92.1	70.1	72.8	70.9	54.9
OmniLMM 12B	81.5	94.4	60.2	50.8	63.2	35.5
MiniCPM-V-2 2.8B	82.1	87.4	63.3	56.4	61.9	37.8

Table 5: Zero-shot-CoT performance of selected models. Compared to True/False prompting, improvements are shown in blue, while declines are shown in red.

As shown in Table 5, the impact of CoT varies depending on the model type and the scores measured. CoT produces significant performance improvements in certain large models, particularly in S_{LP} and LP scores. However, in other instances, the advantages of CoT are negligible, and in some cases, CoT even hinders performance. Additionally, we found that smaller models, such as OminLMM and MiniCPM-V-2, struggled to effectively follow CoT instructions; they generated final answer before the reasoning steps. For these reasons, we adopted an evaluation setting that forces responses to be limited to either "True" or "False."

Language Priors and Model Scale The tendency for the language prior to be inversely proportional to the scale of backbone LLMs may appear counterintuitive (i.e., LLaVA in Table 3). We have not identified the precise cause of this trend. One

possible explanation is that larger pre-trained models are less prone to overfitting to the dataset during the visual instruction tuning process, thereby better maintaining their ability to attend to visual information.

In the experiments, we employ models with various scales of image encoders (ranging from approximately 300M to 5B), however, no clear correlation was observed between the language prior and the size of the image encoder.

Diagnosing LVLMs VLind-Bench can diagnose a model's capabilities in multiple aspects and components, providing clues on where to focus for comprehensive improvements. For instance, a low S_{LP} score suggests that enhancements should be in the vision-language training aspect, while a low S_{CK} score indicates that improvements should focus on the knowledge aspect of the backbone LLM. In the case of the former, utilizing the RLHF-V techniques can significantly reduce the model's reliance on language priors, as observed in Section 5.

7 Conclusion

In this work, we proposed VLind-Bench, a benchmark designed to precisely measure language priors in LVLMs. We evaluated several LVLMs using this benchmark and analyzed the results, finding that the reliance on language priors is inversely proportional to the model scale. Additionally, the RLHF-V technique turned out to significantly aid in reducing such reliance. As demonstrated with VLind-Bench, we endorse a pipelined evaluation paradigm for the general construction of benchmarks to disentangle the specific abilities intended for measurement.

Limitations

Although VLind-Bench minimized potential confounding factors in assessing language priors, there may still be unconsidered factors that contribute to the benchmark scores. VLind-Bench used only a single fixed prompt for evaluation, but recent studies have shown that LLMs and LVLMs respond sensitively to even small changes in such prompts (Zhou et al., 2024b; Lee et al., 2024).

Additionally, the CBs in our benchmark does not necessarily need to receive both text and image as input to check the commonsense bias. Such design choice is mostly due to a lack of established practices for feeding text-only inputs to LVLMs. As alternatives to I_{cf}, we conducted experiments using a plain single-color image or rendered text prompts as visual input (refer to Appendix B); however, none of these approaches works – these kinds of images can be considered out-of-distribution samples, and some proprietary models output error messages for these inputs. Exploring more established methods for text-only inputs in LVLMs falls outside the scope of our paper, but further research in this area is necessary both from a practical perspective and for a deeper understanding of how individual components of LVLMs operate.

Although VLind-Bench addressed various aspects of language priors and commonsense biases, its limitation is that it did not cover social bias (Hall et al., 2023; Lee et al., 2023) or toxicity (Kim et al., 2024; Koh et al., 2024).

Finally, we did not train the LVLMs with the data we constructed. While our primary goal in Section 4 was to generate data for a benchmarking purpose, we can also use this process to generate training data automatically. Training LVLMs with such dataset could help mitigate reliance on language priors, but we leave this as future work.

Acknowledgments

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (RS-2024-00348233), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University) & RS-2021-II212068, Artificial Intelligence Innovation Hub (Artificial Intelligence Institute, Seoul National University)], and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024. K. Jung is with ASRI, Seoul National University, Korea.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren,

Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. ArXiv: 2405.17220.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. 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 (ICCV)*.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. 2019. Monet: Unsupervised scene decomposition and representation. *Preprint*, arXiv:1901.11390. ArXiv: 1901.11390.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805. ArXiv: 2312.11805.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Minbeom Kim, Jahyun Koo, Hwanhee Lee, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 2024. LifeTox: Unveiling implicit toxicity in life advice. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 688–698, Mexico City, Mexico. Association for Computational Linguistics.

Hyukhun Koh, Dohyung Kim, Minwoo Lee, and Kyomin Jung. 2024. Can LLMs recognize toxicity? a structured investigation framework and toxicity metric. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6092–6114, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2024. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. *arXiv preprint arXiv:2410.20774*.

Minwoo Lee, Hyukhun Koh, Kang-il Lee, Dongdong Zhang, Minsung Kim, and Kyomin Jung. 2023. Target-agnostic gender-aware contrastive learning for mitigating bias in multilingual machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16825–16839, Singapore. Association for Computational Linguistics.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744. ArXiv: 2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2024. Probing visual language priors in vlms. *arXiv preprint arXiv:2501.00569*.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2024. Mitigating hallucinations in large vision-language models via summary-guided decoding. *arXiv preprint arXiv:2410.13321*.

Ella Neeman, Roee Aharoni, Or Honovich, Leshem Choshen, Idan Szpektor, and Omri Abend. 2023. DisentQA: Disentangling parametric and contextual knowledge with counterfactual question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10056–10070, Toronto, Canada. Association for Computational Linguistics.

NousResearch. 2023. Nous hermes 2 - yi-34b.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card.

OpenAI. 2023c. Improving image generation with better captions.

OpenAI. 2024. Hello gpt-4o.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *Preprint*, arXiv:2306.14824. ArXiv: 2306.14824.

Vicuna Team. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944. ArXiv: 2310.16944.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023. Evaluation and analysis of hallucination in large vision-language models. *Preprint*, arXiv:2308.15126. ArXiv: 2308.15126.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178. ArXiv: 2304.14178.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *Preprint*, arXiv:2312.00849. ArXiv: 2312.00849.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *Preprint*, arXiv:2405.17220. ArXiv: 2405.17220.

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics.

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

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024a. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*.

Yue Zhou, Yada Zhu, Diego Antognini, Yoon Kim, and Yang Zhang. 2024b. Paraphrase and solve: Exploring and exploiting the impact of surface form on mathematical reasoning in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2793–2804, Mexico City, Mexico. Association for Computational Linguistics.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.

A Human Verification and Model Prompt Details

Criteria for Instance Triple Verification The reviewers are provided with the context, the true statement, and the false statement (which was defined as *instance triple* in the Section 4). For each instance triple, the reviewers are given two options: Accept and Reject. The appropriateness is verified based on the following criteria.

1. Decisions are made based solely on the text without considering image generation.
2. If a true (false) statement is not clearly true (false), it should be rejected.
3. If the context is not counterfactual, it should be rejected.
4. Even if a true (false) statement is indeed true (false), it should be rejected if it does not address the counterfactual aspect of the context.
5. If the truth values of statements cannot be inferred from the context, it should be rejected.
6. Annotators may use internet searches to determine the appropriateness of the context and statement.

Criteria for Image Verification The reviewers are provided with the context, the true statement, the false statement, and the generated image. For each image, the reviewers are given two options: Accept and Reject. The appropriateness is verified based on the following criteria.

1. If a true (false) statement is not clearly true (false), it should be rejected.
2. Accept the image if it is sufficient to determine the truth values of the statements, even if the image does not precisely depict the context.
3. Reject if the generated image is of significantly poor quality.
4. Annotators may use internet searches to determine the appropriateness of the image.

Each instance triple or image was reviewed by a total of three reviewers. Only those instance triples or images that were accepted by at least two reviewers were included in our benchmark.

Prompt Template for Instance Triple Generation We used the following prompt template for instance triple generation. To facilitate understanding of the reader, the template is filled with examples of the concept "location," with the filled-in sections indicated in *italics*.

> Given a concept, create related counterfactual situation (context) which can be described with an image. Also generate two statements with different truth values for each situation. Make only clear statements so that there is no room for vague or different truth value of the statement depending on the point of view. For example, through the concept of *"location"*, we can create a counterfactual situation such as *"A variety of marine life lives in the city built underwater."* and describe it with an image of *a underwater city*. And then we can make two statements, *"The city's buildings are surrounded by marine life."* and *"The city has human residents."*, which is true and false under given counterfactual situation, respectively. List 50 context and statement pairs for the concept of *"location."* Output the results using the following json template.
>
> *[{"id": 1, "context": "A ship is located in the middle of a large city.", "true_statement": "The ship is surrounded by buildings.", "false_statement": "The ship is in the ocean."}, {"id": 2, "context": "A glacier is found in a tropical jungle.", "true_statement": "The glacier coexists with tropical trees.", "false_statement": "The glacier is in the polar region."}, ...]*

Prompt Template for Generating Nouns for VPs As described in Section 4, we employed GPT-4 to extract one key noun from T_{cf} and generate one arbitrary noun not present in T_{cf}, to construct s_{exist} and s_{nil}. To ensure appropriateness, two instances of each noun were initially generated, after which a manual selection process was conducted to choose the better option between the two.

We used the following prompt template for generating nouns for the VPs.

> Extract nouns from the following context. If there are more than two nouns, pick the two

most important nouns. Also generate two random nouns that are not included in the context. Here are some examples.

Context: Wombats burrow in the frozen tundra, their tunnels creating intricate networks under the snow. {"nouns": ["wombat", "tunnel"], "non-existent_nouns": ["zebra", "closet"]}

Context: The jellybean is heavier than the digital piano. {"nouns": ["jellybean", "piano"], "non-existent_nouns": ["car", "oven"]}

Context: *Context*

B Experiments Using a Plain White Image and Rendered Text Prompts

As discussed in Section 6, we conducted experiments using a plain white image and rendered text prompts as visual inputs instead of I_{fact} and I_{cf} in CK and CB. When employing the plain white image, we replaced all images in the CK and CB inputs with a plain white image. In the case of using rendered text prompts, we substituted all CK and CB input images with images that had the content of the textual prompts rendered in black text on a white background.

Table 6: Experimental results on VLind-Bench using various visual inputs.

	Pipeline Score				Accuracy	
Models	S_{CK}	S_{VP}	S_{CB}	S_{LP}	CB	LP
I_{fact} / I_{cf} as visual input						
GPT-4o	93.0	96.0	96.8	89.8	97.0	89.4
LLaVA-NEXT 72B	94.4	95.7	76.1	58.6	75.5	46.7
OmniLMM 12B	88.1	97.7	78.6	81.4	79.5	66.4
plain white image as visual input						
GPT-4o	85.1	96.0	95.7	88.4	96.4	89.4
LLaVA-NEXT 72B	88.4	95.7	74.9	54.8	74.2	46.7
OmniLMM 12B	79.1	97.7	72.4	81.0	72.8	66.4
rendered text prompts as visual input						
GPT-4o	86.1	96.0	96.5	88.5	97.0	89.4
LLaVA-NEXT 72B	89.1	95.7	70.6	54.2	72.8	46.7
OmniLMM 12B	74.2	97.7	65.2	77.4	68.2	66.4

Table 4 presents the results of this experiment, showing a notable performance decline, particularly in the CK. This performance decline can be attributed to the absence of information that was present in the original images. Additionally, both plain white image and rendered text prompts can be considered out-of-distribution inputs (OOD), leading to unstable performance.

C Model Performance by Image Style

Here, we observed how performance varies across different image styles. As mentioned in Section 4, we generated images in photorealistic, illustration, and cartoon styles.

Table 7: Experimental results on VLind-Bench with varying image styles.

	Pipeline Score				Accuracy	
Models	S_{CK}	S_{VP}	S_{CB}	S_{LP}	CB	LP
photorealistic						
GPT-4o	93.1	96.2	97.1	92.3	97.3	91.6
LLaVA-NEXT 72B	94.6	95.8	77.2	65.0	76.5	52.4
OmniLMM 12B	88.8	97.7	81.8	82.8	83.1	70.5
illustration						
GPT-4o	92.7	95.4	97.5	90.1	97.7	90.0
LLaVA-NEXT 72B	94.3	96.2	78.5	59.1	77.8	47.3
OmniLMM 12B	88.5	98.1	81.4	80.4	82.4	67.7
cartoon						
GPT-4o	94.1	96.7	97.2	91.9	97.4	91.5
LLaVA-NEXT 72B	94.8	95.5	78.8	58.2	78.8	48.0
OmniLMM 12B	87.7	97.8	82.2	82.0	82.5	68.4

Table 5 shows that the performance across these styles in the CK, VP, and CB did not vary significantly. A notable variation in performance was observed only in LP, where the photorealistic style yielded better results compared to the other two styles. This could be due to the model's assessment that images in the illustration or cartoon styles lack realism compared to photorealistic images, leading it to generate responses that align more closely with common sense.

D Model Performance on Handcrafted Evaluation Set

To verify the validity of automatically generated text and images, we created a small handcrafted evaluation set and assessed the performance of several models, comparing it with their performance on the original VLind-Bench. All the text in the handcrafted evaluation set was written by humans, with three "triples" for each concept. For each triple, we generated three counterfactual images and one factual image. It was extremely challenging to find real images that depict counterfactual situations, and even if such images were found online, there was no way to verify whether they were outputs from generative models. To eliminate any potential advantage that OpenAI models might have from using DALL-E 3-generated images, we generated all the images using Stable Diffusion and Adobe Firefly, incorporating various styles such

as photorealistic, illustration, and cartoon. This handcrafted evaluation set ultimately consists of 33 triples and 132 images, and the performance on this set is as follows.

Table 8: Experimental results on handcrafted evaluation set.

	Pipeline Score				Accuracy	
Models	S_{CK}	S_{VP}	S_{CB}	S_{LP}	CB	LP
GPT-4o	90.9	90.9	100.0	93.8	100.0	93.9
GPT-4V	90.9	81.8	90.0	75.8	87.9	72.7
LLaVA-NEXT 72B	87.9	97.0	79.3	60.9	78.8	47.5
OmniLMM 12B	81.8	97.0	85.2	76.8	87.9	70.7
MiniCPM-V-2 2.8B	63.6	97.0	66.7	50.0	57.6	39.4

As shown in Table 8, the performance on the original VLind-Bench and the gold evaluation set does not differ significantly (refer to Table 3 for the original VLind-Bench scores).

E Data Samples

Here, we provide some data samples for each concept (next page). For the notations, please refer to the section 3.

F Information About Use Of AI Assistants

In writing this paper, we utilized ChatGPT [2] for paraphrasing.

[2]https://chatgpt.com/

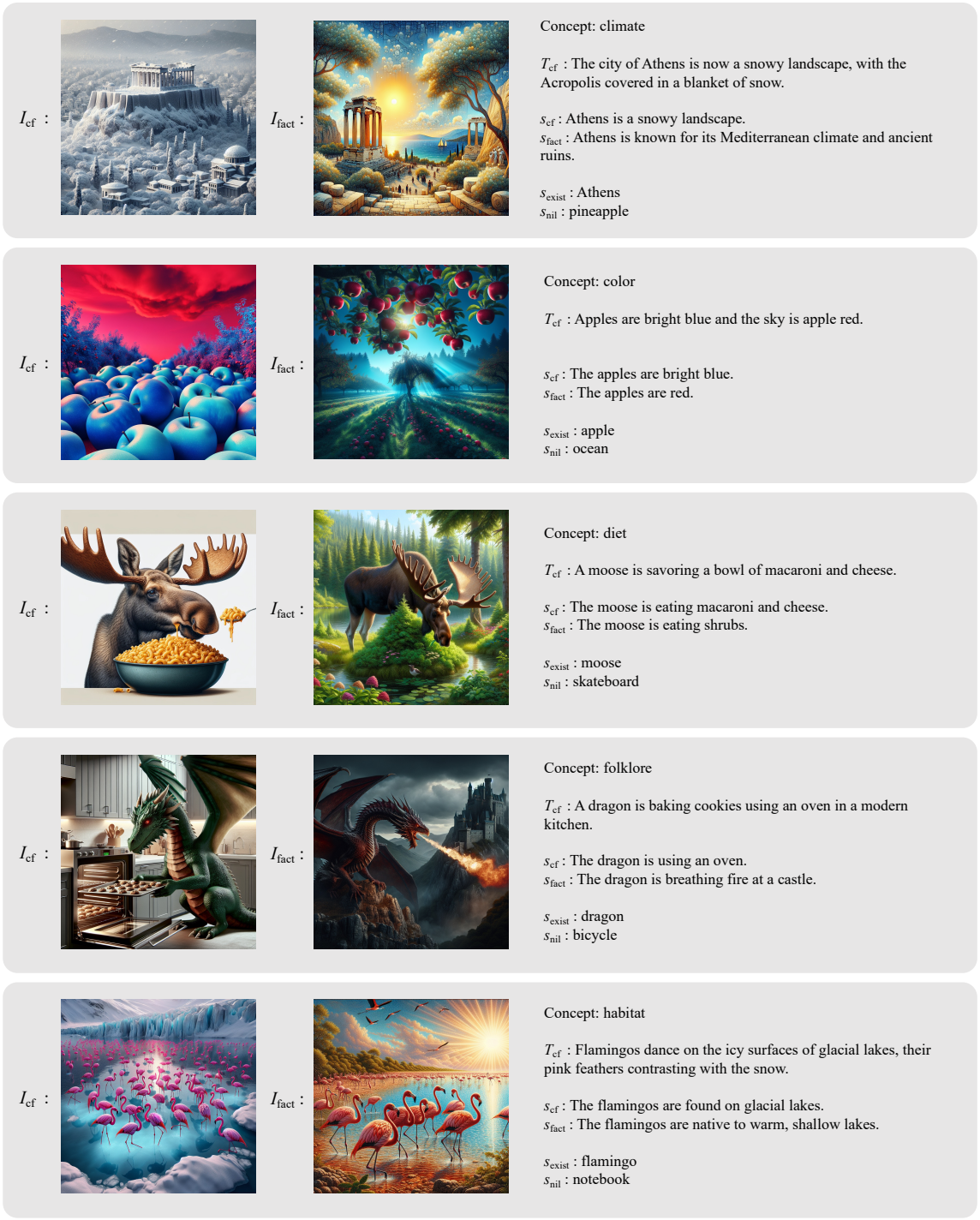


Figure 2: Data samples for concept of climate, color, diet, folklore, and habitat.

I_{cf} :  I_{fact} : 

Concept: history

T_{cf} : Viking settlements thriving in a rainforest.

s_{cf} : Vikings established colonies in a rainforest.
s_{fact} : Vikings did not colonize a rainforest.

s_{exist} : rainforest
s_{nil} : laptop

I_{cf} :  I_{fact} : 

Concept: landmark

T_{cf} : The Sphinx in Egypt is depicted with a lion's head instead of a human head.

s_{cf} : The Sphinx has a lion's head.
s_{fact} : The Sphinx has a human head.

s_{exist} : Sphinx
s_{nil} : bicycle

I_{cf} :  I_{fact} : 

Concept: location

T_{cf} : A ski resort is operating on a tropical island.

s_{cf} : Skiers are going down slopes next to palm trees.
s_{fact} : The ski resort is located in a cold, mountainous region.

s_{exist} : ski resort
s_{nil} : lighthouse

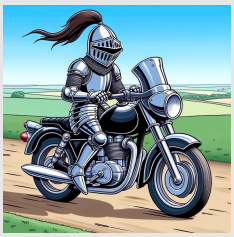
I_{cf} :  I_{fact} : 

Concept: size

T_{cf} : A person is playing with a pet ant, with the ant being as big as a dog.

s_{cf} : The ant is bigger than the person's foot.
s_{fact} : The person's foot is bigger than the ant.

s_{exist} : person
s_{nil} : waterfall

I_{cf} :  I_{fact} : 

Concept: time

T_{cf} : A knight in armor is seen riding a motorcycle instead of a horse.

s_{cf} : Knights rode motorcycles.
s_{fact} : Motorcycles were not available during the time of knights.

s_{exist} : knight
s_{nil} : waterfall

I_{cf} :  I_{fact} : 

Concept: weight

T_{cf} : A gummy bear is making a metal filing cabinet rise when placed on a seesaw.

s_{cf} : The gummy bear is heavier than the metal filing cabinet.
s_{fact} : The metal filing cabinet is heavier than the gummy bear.

s_{exist} : gummy bear
s_{nil} : mountain

Figure 3: Data samples for concept of history, landmark, location, size, time, and weight.