NaturalBench: Evaluating Vision-Language Models on Natural Adversarial Samples

Supplementary Material

Outline

This document supplements the main paper with detailed results. Below is the outline:

- **Section A** details the collection process of NaturalBench.
- **Section B** details VQA and image-text retrieval performance on NaturalBench.
- **Section C** provides skill definitions and analyzes model performance by skills.
- **Section D** reports other debiasing techniques on NaturalBench.

A   Collection Details

We provide further details on the collection pipeline.

Step 1: Collecting pairs of image-text samples. We collect pairs of image-text samples by finding mismatches of discriminative VLMs like CLIP. Recall that VLMs like CLIP [65] compute a similarity score $S(\mathbf{i}, \mathbf{t}) \in \mathbb{R}$, with higher scores indicating greater similarity between the image $\mathbf{i}$ and text caption $\mathbf{t}$. For a pair of image-text samples $\{(\mathbf{i}_0, \mathbf{t}_0), (\mathbf{i}_1, \mathbf{t}_1)\}$, a mismatch occurs when:

$$[S(\mathbf{i}_0, \mathbf{t}_0) < S(\mathbf{i}_0, \mathbf{t}_1)] \quad \text{or} \quad [S(\mathbf{i}_0, \mathbf{t}_0) < S(\mathbf{i}_1, \mathbf{t}_0)] \quad \text{or} \quad [S(\mathbf{i}_1, \mathbf{t}_1) < S(\mathbf{i}_0, \mathbf{t}_1)] \quad \text{or} \quad [S(\mathbf{i}_1, \mathbf{t}_1) < S(\mathbf{i}_1, \mathbf{t}_0)] \qquad (3)$$

Importantly, this adversarial procedure efficiently pairs similar image-text samples for two purposes. First, these image-text pairs already form an image-text retrieval task that can be evaluated using Winoground's [71] evaluation protocols (after removing pairs where one caption can describe both images). We term this benchmark **NaturalBench-Retrieval** and report the performance of CLIP and SigLIP in Table 7. Next, by considering both images and captions, we can pair samples that are semantically similar but not necessarily visually similar. This contrasts with MMVP [73] which only pairs visually similar images close in DINO's feature space.

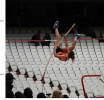
Implementation of step 1. For Flickr30K [63], we retrieve pairs mismatched by both OpenCLIP (LAION400M-ViT-L14) [29] and BLIP-2 (ViT-L) [39]. For DOCCI [59], we use both longCLIP-B and longCLIP-L. However, since DOCCI's captions are still too long to process, we use ChatGPT to shorten them to below 230 characters per caption. We believe future advances in long-context CLIP will streamline this process. Lastly, for XM3600, we use NLLB-CLIP [75] to process the Chinese and Hindi captions.

Step 2: Generating questions and answers. We use ChatGPT to generate questions that yield different answers for two images using their textual captions. We now show the actual prompts we send to ChatGPT.

Default instruction for GPT-4. In practice, we use the below prompt to ask GPT-4 to directly output a JSON dictionary for easier processing:

> I will present two captions for two images. Please help me generate two questions that highlight the differences between the captions. The first question should result in a 'Yes' answer for **Caption 1** and a 'No' for **Caption 2**. The second question should result in a 'No' answer for **Caption 1** and a 'Yes' for **Caption 2**.
> **Caption 1:** {$\mathbf{t}_0$}
> **Caption 2:** {$\mathbf{t}_1$}
> Please response in JSON format with question indices as the keys, starting from 0 and question-answer pairs {{"Question":...,"Caption1 Answer":...,"Caption2 Answer":...}} as the values.

Instructions for generating Chinese and Hindi QA pairs. We can simply ask GPT-4 to generate questions and answers in Chinese and Hindi:

16

I will present two captions for two images. Please help me generate two questions in Chinese / Hindi that highlight the differences between the captions. The first question should result in a 'Yes' answer for **Caption 1** and a 'No' for **Caption 2**. The second question should result in a 'No' answer for **Caption 1** and a 'Yes' for **Caption 2**.
Caption 1: {t_0}
Caption 2: {t_1}
Please response in JSON format with question indices as the keys, starting from 0 and question-answer pairs {{"Question":...,"Caption1 Answer":...,"Caption2 Answer":...}} as the values.

Instructions for generating multiple-choice QA pairs. We ask ChatGPT to generate multiple-choice questions using the below prompt:

I will present two captions for two images. Please help me generate two multiple-choice questions that highlight the differences between the captions. Each question should have options A and B. For the first question, option A corresponds to **Caption 1** and option B corresponds to **Caption 2**. For the second question, option A corresponds to **Caption 2** and option B corresponds to **Caption 1**.
Caption 1: {t_0}
Caption 2: {t_1}
Please response in JSON format with question indices as the keys, starting from 0 and question-answer pairs {{"Question":...,"Caption1 Answer":...,"Caption2 Answer":...}} as the values.

We engage two human annotators to select from the two candidate answers and "Unanswerable" [22] for all generated QA pairs, retaining a sample only if both annotators agree on the correct answer. In total, we spend around 500 annotator hours to collect all samples at 14 dollars per hour. For the Chinese and Hindi subsets, the authors (who are native speakers of these languages) manually examine all the questions.

Additional examples. Figure 7 provides additional examples of NaturalBench.

Figure 7: **More NaturalBench examples.**

B   NaturalBench Performance

We report model performance on different subsets of NaturalBench.

Performance on different subsets. Table 4 reports **G-Acc** on subsets of NaturalBench.

Table 4: **Performance on different subsets of NaturalBench.** We report the **G-Acc** performance of 53 leading VLMs on subsets of NaturalBench.

Model	Image Encoder	Language Model	NaturalBench Performance				
			Flickr-YN	Flickr-MCQ	DOCCI-YN	DOCCI-MCQ	Overall
Human Performance	–	–	91.5	92.0	92.2	93.9	92.1
Random Chance	–	–	6.3	6.3	6.3	6.3	6.3
Open-source Models							
BLIP-2 [39]	EVA-G	FlanT5-3B	2.7	0.8	2.3	0.5	2.1
		FlanT5-11B	6.1	3.2	12.1	1.0	7.7
InstructBLIP [9]	EVA-G	Vicuna-7B	2.9	0.4	6.8	0.5	4.0
		Vicuna-13B	7.0	0.4	14.8	5.0	9.2
		FlanT5-3B	9.6	1.2	15.1	0.5	9.8
		FlanT5-11B	12.6	2.8	18.6	2.5	12.7
Otter [33]	CLIP-L-14	MPT-7B	3.7	4.0	4.5	1.5	3.8
LlaMA-Adapter-v2.1 [19]	CLIP-L-14	LlaMA2-7B	4.2	1.2	6.6	0.5	4.4
CogVLM-Agent-VQA [25]	EVA2-E	Vicuna-7B	12.2	2.8	13.6	0.5	10.3
DeepSeek-VL-1.3B-Chat [51]	SigLIP-L & SAM-B	DeepSeek-LLM-1B	7.8	3.6	15.5	17.0	11.5
LLaVA-1.5 [47]	CLIP-L-14	Vicuna-7B	9.1	14.8	14.1	16.5	12.7
		Vicuna-13B	9.1	21.2	15.1	24.0	14.8
ShareGPT4V	CLIP-L-14	Vicuna-7B	10.0	13.2	12.9	18.5	12.5
		Vicuna-13B	9.5	19.6	15.6	23.5	14.9
CogVLM-Chat [77]	EVA2-E	Vicuna-7B	14.6	15.6	14.5	7.5	13.9
InternLM-XC-V1 [87]	EVA-G	InternLM-7B	11.5	16.8	15.2	28.0	15.5
InternLM-XC-V2-1.8B [15]	CLIP-L-14	InternLM2-1.8B	12.0	25.6	15.1	26.5	16.6
Qwen-VL-Chat [3]	CLIP-G-16	Qwen-7B	16.0	16.8	16.9	21.5	17.1
Phi-3-Vision [1]	CLIP-L-14	Phi-3-Mini	15.4	17.6	15.3	30.0	17.2
mPLUG-Owl2 [81]	CLIP-L-14	LlaMA2-7B	14.0	20.0	17.3	25.5	17.4
Bunny [23]	SigLIP-SO	Phi-2-2.7B	12.0	16.8	18.9	30.0	17.4
mPLUG-Owl2.1 [81]	CLIP-L-14	Qwen-7B	12.3	20.0	17.4	36.0	17.9
Monkey-10B-chat [41]	OpenCLIP-BigG	Qwen-7B	17.1	12.0	19.5	24.0	18.2
LLaVA-NeXT [48]	CLIP-L-14	Vicuna-7B	12.5	17.6	14.5	22.0	15.0
		Mistral-7B	13.7	21.6	14.6	24.5	16.3
		Vicuna-13B	15.7	22.8	19.0	26.5	19.2
		Nous-Hermes-2-Yi-34B	16.2	32.0	20.8	40.0	22.7
DeepSeek-VL-7B-Chat [51]	SigLIP-L & SAM-B	DeepSeek-LLM-7B	13.8	18.8	21.6	28.5	19.3
BLIP-3 (XGen-MM) [79]	CLIP-H-14	Phi-3-Mini	13.7	19.2	21.6	30.5	19.5
InternVL-Chat-V1.1 [6]	InternViT-6B	LlaMA2-13B	19.7	21.6	16.5	36.0	20.3
InternVL-Chat-V1.5 [7]	InternViT-6B	InternLM2-Chat-20B	22.5	32.8	17.4	35.5	23.1
InternVL-Chat-V1.2-Plus [6]	InternViT-6B	Nous-Hermes-2-Yi-34B	26.5	31.2	17.0	29.5	23.4
InternVL2-8B [8]	InternViT-300M	InternLM2.5-7B-Chat	20.7	34.8	19.5	35.0	23.5
Cambrian-1 [72]	SigLIP-S-14 & CLIP-L-14	Llama-3-8B	16.2	15.6	24.6	14.0	19.4
	DINOv2-g &	Vicuna-13B	19.6	30.8	26.1	35.5	25.5
	CLIP-ConvNeXT-XXL	Nous-Hermes-2-Yi-34B	23.8	35.2	23.7	36.5	26.6
InternLM-XC2-4KHD-7B [13]	CLIP-L-14	InternLM2-7B	22.8	33.2	24.3	34.0	25.9
InternLM-XC2-7B [15]	CLIP-L-14	InternLM2-7B	25.4	38.4	21.8	34.0	26.5
InternVL-Chat-V1.2 [6]	InternViT-6B	Nous-Hermes-2-Yi-34B	21.8	34.4	24.5	41.0	26.6
InternVL2-26B [8]	InternViT-6B	InternLM2-Chat-20B	26.6	**40.4**	22.1	37.5	27.7
LLaVA-OneVision [37]	SigLIP-S-14	Qwen2-0.5B	12.0	14.4	18.6	17.5	15.6
		Qwen2-7B	27.0	28.0	26.0	41.5	28.8
Llama3.2-Vision [17]	ViT-H-14	Llama-3.1-8B	16.2	29.2	30.0	45.5	26.8
		Llama-3.1-70B	23.7	37.2	24.8	**53.5**	29.1
Molmo [10]	CLIP-L-14	OLMoE-1B-7B	10.8	15.2	14.3	29.0	14.7
		OLMo-7B	14.6	24.0	20.6	37.0	20.7
		Qwen2-7B	20.6	31.2	25.8	44.5	26.7
		Qwen2-72B	23.5	38.4	25.3	52.5	29.3
Qwen2-VL [76]	CLIP-L-14	Qwen2-1.5B	17.4	22.8	25.6	35.0	23.4
		Qwen2-7B	18.8	28.4	32.9	48.5	29.1
		Qwen2-72B	28.2	36.0	**40.5**	52.0	36.9
Closed-source Models							
GPT-4Vision	–	GPT-4	22.8	25.2	26.9	36.0	26.2
GPT-4o	–	GPT-4	**37.5**	**40.4**	39.0	48.0	**39.6**

Performance on NaturalBench-Hindi and NaturalBench-Chinese. Table 5 reports the performance on the multilingual subsets of NaturalBench, evaluating only the models that claim to have multilingual capabilities. We also report the performance of these datasets after using ChatGPT to translate the questions and answers into English. This shows that most models are still better at solving English VQA tasks.

Ablation on samples generated by different methods. Table 6 reports **G-Acc** on two types of generated VQA samples: (1) **Flickr-Adversarial**, generated by sending caption pairs to GPT-4, (2)

18

Table 5: **Performance on NaturalBench-Chinese and NaturalBench-Hindi.** We report **G-Acc** for each dataset, evaluating only models with claimed multilingual capabilities. For both datasets, we also provide G-Acc after translating the original Chinese or Hindi questions into English. This simple translation often boosts performance, except for top models like InternVL-Chat-V1.2-Plus and GPT-4o, which seem extensively trained in Chinese. NaturalBench-Hindi remains particularly challenging for open-source models.

Model	NaturalBench-Chinese		NaturalBench-Hindi	
	Chinese	English	Hindi	English
Random Chance	6.3	6.3	6.3	6.3
Open-source Models				
DeepSeek-VL-7B-Chat	10.9	**28.4**	0.6	**29.0**
InternVL-Chat-V1.2-Plus	**34.6**	33.4	11.5	**36.2**
InternLM-XC2-7B	32.5	**34.6**	15.9	**35.6**
Closed-source Models				
GPT-4o	**41.2**	38.7	40.3	**40.9**

Table 6: **Ablation on different collection methods.** We report **G-Acc** on datasets generated by different collection methods from Flickr30K. Our adversarial procedure results in a much more challenging dataset. Note that Flickr-Adversarial is the combination of Flickr-YN and Flickr-MCQ.

Model	Model Performance (G-Acc)	
	Flickr-Adversarial	Flickr-Random
Random Chance	6.3	6.3
Open-source Models		
DeepSeek-VL-7B-Chat	15.2	80.7
BLIP-3(XGen-MM)	15.2	69.0
LLaVA-NeXT (Mistral-7B)	15.9	86.0
Phi-3-Vision	16.0	75.0
InternVL-Chat-V1.2-Plus	27.8	83.0
InternLM-XC2-7B	29.0	84.5
Closed-source Models		
GPT-4o	38.3	72.5

Flickr-Random, generated by sending caption pairs of *randomly matched* image-text samples to GPT-4. The results confirm that it is crucial to use discriminative VLMs to first search for confounding pairs of image-text samples.

Performance on NaturalBench-Retrieval. Table 7 reports model performance on NaturalBench-Retrieval. We only use Flickr image-text samples to construct this benchmark. We adopt the evaluation metrics proposed by Winoground [71].

19

Table 7: **Image-text retrieval performance on NaturalBench-Retrieval.** We evaluate CLIP and SigLIP models on the human-verified 1,200 paired (image, text) samples from NaturalBench-Flickr. We follow Winoground [71] to report text score, image score, and group score, with higher numbers indicating better performance for all metrics. We exclude the CLIP (LAION400M-ViT-L14) model used to collect these adversarial pairs. Overall, NaturalBench-Retrieval poses a significant challenge to leading discriminative models.

Method	Source	Model	Data Size	Model Size (M)	Retrieval Performance		
					Group	Image	Text
Random	–	–	–	–	16.67	25.00	25.00
CLIP [65]	OpenAI	RN50	400M	102	12.22	32.60	36.76
		RN101		120	13.61	35.04	33.33
		ViT-B-32		151	15.89	36.43	36.92
		RN50x4		178	14.75	37.49	36.27
		RN50x16		291	24.61	44.01	43.93
		ViT-L-14		428	23.15	44.99	41.81
		RN50x64		623	26.24	46.21	47.35
	LAION	roberta-ViT-B-32	2B	212	16.22	39.36	38.79
		ViT-H-14		986	24.04	49.31	48.82
		ViT-g-14		1367	21.35	46.21	46.54
		ViT-bigG-14		2540	21.04	44.49	43.69
		xlm-roberta-base-ViT-B-32	5B	366	16.79	37.49	40.91
		xlm-roberta-large-ViT-H-14		1193	22.82	47.35	47.51
	DataComp	small: ViT-B-32	13M	151	12.06	22.90	21.19
		medium: ViT-B-32	128M	151	16.95	28.28	33.01
		large: ViT-B-16	1B	150	16.71	36.43	35.86
		xlarge: ViT-L-14	13B	428	21.84	44.01	45.72
SigLIP [85]	WebLI (English portion)	ViT-B	13B	172	24.29	48.57	49.06
		ViT-L		430	31.21	54.93	54.44
		ViT-SOViT		800	**42.14**	**62.67**	**63.90**

20

C Skill Analysis

We now provide the skill definitions and report model performance by each skill tag.

Skill definitions and examples. Table 8 provides definitions to the skills in NaturalBench.

Skill analysis. Table 9 reports **Q-Acc** performance (awarding one point if the model answers both images correctly for each question) on **Object** and **Attribute** tags. Table 10 reports **Q-Acc** performance on **Relation** and **Reasoning** tags.

Additional examples. We provide additional tagging examples in Figure 8. We will release these tags for more fine-grained analysis, such as evaluating models on combinations of skills.

Figure 8: **More NaturalBench examples with skill tags.**

21

Table 8: **Skill definitions.**

Skill Type	Definition	Examples
Object	Basic entities within an image, including animals, humans, food, buildings, natural elements (nature), vehicles, common items, and others.	*Is there a **car** parked near the path? Is there a **person** in this image? Is there a **referee** behind the **table**? Is the **dog** fully submerged in the **water** except for its **head**? Is the **water body** filled with visible **rocks** and emanating **ripples**?*
Attribute	Visual properties of entities, including emotion, shape, size, color, state, activity, gender, and abstract attributes (e.g., helpful, lucky).	*Is anyone in the picture **sad** or **scared**? Is the woman extremely **surprised**? Is the woman alone behind a **glass** partition? Is the man wearing **brown**? Is the man wearing a **red and white striped** apron? Is the **old** man in the image wearing **reflective** safety jackets?*
Spatial Relation	Physical arrangements of multiple entities relative to each other [46], including proximity (e.g., near, far), topological (e.g., at, on, in, with, surround, between, inside, outside) , projective (e.g., left of, right of, under, in front of, below), orientation and direction (e.g., facing, towards, across, away from).	*Is there a referee **behind** the table? Is the dog looking **up** at the sky? Is there only one person **in** the canoe? Is there a group of people standing **outside** the gates? Is the man in the image looking **at** the object to his **left**? Is the smiling woman standing **next to** the bus?*
Action Relation	Action interactions between entities, e.g., pushing, kissing, hugging, hitting, helping, and so on.	*Is there a person **holding** a water bottle? Is the black dog **biting** a stick? Is anyone **using** an umbrella? Is the man **holding** a red pen? Is the dog **chasing after** a toy outdoors? Is the person **jumping directly off** a building without any equipment?*
Part Relation	Part-whole relationships between entities – one entity is a component of another, such as body part, clothing, and accessories.	*Is there a person wearing orange and yellow **shirt** and **jacket**? Is anyone wearing yellow and orange **safety vests**? Is the woman in the black **dress** wearing **gloves**? Is a player using **his back** to play the ball? Is the **boy's tongue** sticking out?*
Counting	Determining the quantity, size, or volume of entities, e.g., objects, attribute-object pairs, and object-relation-object triplets.	*Are there **four** people in the image? Does the dog have **two** visible colors? Are there **more than four** performers in the image?*
Differentiation	Differentiating objects within a category by their attributes or relations, such as distinguishing between "old" and "young" people by age, or "the cat on top of the table" versus "the cat under the table" by their spatial relations.	*Does the girl on the left look sad while the girl on the right look happy? Is there a cat sitting on a grey cabinet in front of another cat sitting on the stairs? Is one dog biting the ear of the other dog? Is a man standing behind another man sitting at a desk?*
Comparison	Comparing characteristics like number, attributes, area, or volume between entities.	*Does the scene involve players from three **different** team colors? Does the **tallest** building feature glass windows and side slopes? Is the **older** person following the **younger** one? Are there two dogs that are significantly **different** in size? Is the man wearing the **same** color as the woman in the image?*
Logic	Understanding logical operators. We only consider negation (as indicated by "no", "not", or "without") and universality (as indicated by "every", "all", "each", "both"). Other logical relations such as conjunction (as indicated by "and", "or") are omitted.	*Does the image show **all** men performing the same action? Are **both** people looking in the same direction? Is the bicycle rider performing a trick **without** any audience? Is the main subject **not** wearing shirt and lying down? Is the main activity potentially related to craft **or** construction?*
World Knowledge	Answering based on external commonsense knowledge, including social, symbolic, functional, physical, natural knowledge and so on.	*Is the event related to the Olympics? Is there a vertical depiction of Ramses III in the image? Does the image suggest a relatively informal social gathering? Is a single individual attempting to score regardless of multiple defenders?*

22

Table 9: **Model performance on Object and Attribute.** We report **Q-Acc** on each tag.

Model	Object								Attribute							
	Animal	Human	Food	Building	Nature	Vehicle	Items	Others	Emotion	Shape	Size	Color	State	Abstract	Activity	Gender
BLIP-3(XGen-MM)	18.6	16.2	15.4	20.8	21.7	22.2	21.2	17.6	9.1	19.3	24.1	21.8	20.2	20.4	16.5	14.0
Phi-3-Vision	15.6	17.1	15.4	17.7	15.6	19.0	18.5	16.7	18.2	17.5	19.0	18.9	16.8	15.6	15.2	15.8
DeepSeek-VL-7B-Chat	20.9	16.9	15.4	21.9	22.1	16.7	19.3	19.0	12.1	24.6	21.4	20.8	19.5	16.7	20.1	14.6
LLaVA-NeXT(Mistral-7B)	14.2	16.1	17.3	14.0	13.4	18.1	16.7	15.2	15.2	19.3	14.6	16.3	15.7	14.1	14.4	17.9
InternLM-XC-V2-7B	23.3	28.6	19.2	30.8	23.6	30.6	27.8	29.0	33.3	31.6	30.2	27.8	25.8	23.3	27.0	30.1
InternVL-Chat-V1.2-Plus	23.9	28.0	23.1	20.3	18.5	22.7	25.4	19.7	21.2	17.0	20.0	24.8	22.8	19.3	26.2	30.4
GPT-4o	**35.4**	**39.7**	**44.2**	**40.1**	**41.3**	**38.4**	**42.8**	**38.3**	**39.4**	**42.1**	**40.7**	**39.0**	**41.1**	**38.9**	**35.5**	**43.2**

Table 10: **Model performance on Relation and Reasoning.** We report **Q-Acc** on each tag.

Model	Relation						Reasoning				
	Action	Part	Proximity	Topological	Projective	Orientation	Count	Logic	Differ	Compar	World
BLIP-3(XGen-MM)	18.3	17.4	27.5	22.8	19.6	15.5	20.6	15.9	13.0	20.9	5.3
Phi-3-Vision	16.0	19.5	19.6	17.9	13.9	9.5	16.1	18.5	17.6	13.0	8.5
DeepSeek-VL-7B-Chat	17.5	16.2	29.4	21.4	17.9	14.7	19.6	16.4	11.1	11.3	10.6
LLaVA-NeXT(Mistral-7B)	15.9	18.6	18.6	17.0	16.1	13.8	17.1	21.2	17.6	12.2	9.6
InternLM-XC-V2-7B	27.3	29.3	29.4	27.9	24.4	24.1	30.7	25.9	27.8	27.8	17.0
InternVL-Chat-V1.2-Plus	23.6	28.1	31.4	24.4	19.3	18.1	23.9	26.9	25.0	15.7	12.8
GPT-4o	**39.4**	**43.1**	**40.2**	**41.7**	**38.7**	**35.3**	**39.2**	**42.9**	**38.9**	**37.4**	**35.1**

23

D Debiasing Analysis

In the main paper, we show that debiasing within the image-text pairings significantly improves model performance. Here, we explore debiasing techniques that don't rely on knowing the image-question pairings.

Deterministic evaluation using answer likelihood [44]. Recall that we can perform a scoring-based evaluation strategy using the generative likelihood of each candidate answer (VQAScore [44]) to determine the model's predicted answer. Specifically, given a question q, an image i, and two candidate answers a_0 and a_1, we evaluate:

$$P(a_0|q, i) - P(a_1|q, i) > \tau \qquad (4)$$

where τ is a threshold (default is 0). If this condition (Eq. 4) is met, the model predicts a_0; otherwise, it predicts a_1. Crucially, this formulation has two benefits: (1) it produces deterministic results that are almost consistent with stochastic decoding (see Table 11), and (2) it allows us to adjust $\tau \in [-1, 1]$ for debiasing. Recall that our main paper performs **sample**-level debiasing by optimizing τ within each of the four image-question pairs. Alternatively, we can perform **global**-level debiasing by searching for a single τ that maximizes **G-Acc** across all samples. We also implement the **post-hoc** debiasing technique proposed in [88], which is equivalent to:

$$\frac{P(a_0|q, i)}{P(a_0|q)} - \frac{P(a_1|q, i)}{P(a_1|q)} > 0 \qquad (5)$$

where $P(a|q)$ is estimated by sending no image tokens but just the question tokens to the VLM. Table 11 shows that these alternate techniques still lag behind the performance of sample-level debiasing. We hope NaturalBench can be a useful testbed for bias mitigation techniques for VLMs.

Table 11: **Evaluating debiasing techniques on NaturalBench.** We evaluate debiasing techniques (as detailed in Section 5) that do not require prior knowledge of image-question pairings (unlike sample optimal τ). For comprehensiveness, we report both stochastic decoding and deterministic evaluation using VQAScore, finding consistent results. We observe that the two post-hoc methods – global-optimal τ and Post-Hoc debiasing – perform significantly worse than the (oracle) sample-optimal τ. Global optimal τ shows only slight improvements, while Post-Hoc debiasing even reduces performance in models like Bunny, InterVL-Chat-V1.2, and GPT-4o. This suggests NaturalBench can be a valuable benchmark for testing future debiasing methods.

Model	Stochastic Decoding			Deterministic VQAScore			Post-hoc Debiasing [88]			Global Optimal τ			Sample Optimal τ		
	Q-Acc	I-Acc	G-Acc	Q-Acc	I-Acc	G-Acc	Q-Acc	I-Acc	G-Acc	Q-Acc	I-Acc	G-Acc	Q-Acc	I-Acc	G-Acc
LLaVA-1.5 (Vicuna-7B)	37.7	43.8	12.7	36.7	42.7	12.2	38.2	44.5	13.9	39.9	45.8	14.0	83.4	76.3	44.3
LLaVA-1.5 (Vicuna-13B)	39.6	44.6	14.8	38.6	43.5	14.4	38.5	42.8	14.5	42.8	47.8	16.5	86.2	78.6	49.7
Phi3-Vision	43.4	48.7	17.2	43.6	48.9	17.7	45.1	48.6	19.3	44.7	49.3	18.4	85.7	78.5	50.0
Bunny	42.3	48.4	17.4	42.5	48.5	17.5	38.7	44.9	15.7	43.6	49.5	18.7	85.8	78.6	50.5
LLaVA-NeXT (Vicuna-7B)	42.5	47.6	15.0	42.0	47.1	15.0	44.2	48.9	18.0	43.4	48.5	16.5	86.7	79.6	50.3
LLaVA-NeXT (Mistral-7B)	44.6	49.1	16.3	45.0	49.4	17.0	46.8	51.1	19.6	45.3	49.7	17.4	88.3	81.6	56.0
LLaVA-NeXT (Vicuna-13B)	45.9	49.9	19.2	44.6	48.5	18.2	48.7	52.5	21.5	47.8	52.1	20.4	89.1	82.3	57.2
DeepSeek-VL-7B-Chat	46.0	50.1	19.3	45.8	49.9	19.4	-	-	-	46.4	50.4	19.7	86.6	81.8	54.8
BLIP-3(XGen-MM)	47.0	51.2	19.5	46.8	51.1	19.5	47.8	52.0	22.4	48.7	53.2	21.4	88.6	81.9	55.3
InternVL-Chat-V1.5	52.3	55.9	23.1	52.6	56.0	24.3	55.2	58.4	28.6	52.3	55.6	25.0	92.3	86.1	66.0
InterVL-Chat-V1.2-Plus	52.7	56.2	23.4	52.6	56.3	23.5	55.9	58.6	28.3	53.0	56.1	24.6	92.4	85.5	65.3
InternVL2-8B	50.5	54.5	23.6	50.4	54.3	23.7	52.2	55.9	25.5	50.4	54.3	23.7	88.7	83.2	58.6
InternVL-Chat-V1.2	52.9	56.4	26.6	52.6	56.0	26.2	52.3	54.3	25.8	53.6	56.8	27.2	91.6	86.0	65.8
InternVL2-26B	55.9	58.8	28.1	55.7	58.5	28.2	58.8	61.1	32.0	55.7	58.3	28.5	92.2	87.2	67.7
LLaVA-OneVision (Qwen2-0.5B)	39.8	46.3	15.7	39.1	44.6	14.5	39.1	44.5	15.8	39.2	46.3	16.2	84.6	77.2	47.5
LLaVA-OneVision (Qwen2-7B)	56.2	58.8	28.9	55.4	58.2	28.6	59.1	61.2	33.2	56.1	59.0	28.7	92.1	87.2	67.8
GPT-4o	**64.4**	**66.4**	**39.6**	**65.0**	**67.0**	**40.5**	**61.6**	**63.2**	**37.6**	**64.9**	**67.1**	**40.7**	**94.0**	**90.5**	**75.6**

24

Checklist

1. For all authors...

    (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The main claims are reflected in the paper's contributions and scope. See Section 3 to Section 6.

    (b) Did you describe the limitations of your work? [Yes] See Section 5.

    (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5.

    (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...

    (a) Did you state the full set of assumptions of all theoretical results? [N/A]

    (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments (e.g. for benchmarks)...

    (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See Section 3 for detailed benchmark collection pipeline. We have released the code in our project site.

    (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplement.

    (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

    (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplement.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

    (a) If your work uses existing assets, did you cite the creators? [N/A]

    (b) Did you mention the license of the assets? [N/A]

    (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] The assets used are all publicly sourced, and therefore explicit consent was not required.

    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No] The data used are publicly available datasets that do not contain offensive content and with consent for personally identifiable information.

5. If you used crowdsourcing or conducted research with human subjects...

    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes] See Section 3.

    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [Yes] We pay the participants above the minimum wage with an hourly pay.

25