Are Horses Always Strong and Donkeys Dumb? Animal Bias in Vision Language Models

Mohammad Anas
Dept of Computer Science and Engineering
Jamia Hamdard University
New Delhi, India
mohammadanas@jamiahamdard.ac.in

Mohammad Nadeem
Dept of Computer Science
Aligarh Muslim University
Aligarh, India
mnadeem.cs@amu.ac.in

Shahab Saquib Sohail
School of Computing Science and Engineering
VIT Bhopal University
Sehore, MP, India
shahabsaquibsohail@vitbhopal.ac.in

Erik Cambria
College of Computing and Data Science
Nanyang Technological University
Singapore
cambria@ntu.edu.sg

Amir Hussain
School of Computing Engineering and the Built Environment
Edinburgh Napier University
Scotland, UK
a.hussain@napier.ac.uk

***Abstract*—Vision Language Models (VLMs), such as CLIP, are widely used for various multimodal tasks and offer significant advancements in image-text understanding. However, existing studies have revealed that VLMs inherit biases from their training data which lead to the reinforcement of harmful stereotypes and cultural misrepresentations. In the proposed work, we analyze the presence of biases associated with animals in the CLIP model. We introduce a novel taxonomy, called Animal Bias Taxonomy (ABT), which categorizes stereotyped associations of animals in three categories. We also curated an animal dataset from existing datasets and applied data-cleaning process on it to remove unwanted images. Using ABT, we evaluated the outputs of VLMs on animal dataset when prompted with animal-related stereotyped terms to assess whether CLIP propagates biased associations that align with cultural stereotypes. Our findings reveal that CLIP frequently exhibits skewed cultural interpretations, such as associating owls with wisdom. Our study underscores the necessity of bias evaluation in VLMs and calls for greater transparency and culturally diverse data curation to ensure fair and inclusive AI systems. The code is available at https://github.com/MohammadAnas5/Clip-sAnimalStereotyping**

I. Introduction

Vision language models (VLMs) represent a significant advancement in the field of artificial intelligence, combining the capabilities of both computer vision (CV) and natural language processing (NLP) to enable machines to understand and generate multimodal content [1]–[3]. They are powered by transformer-based architectures and are pre-trained on large-scale datasets that contain image-annotation pairs [4]–[6]. It allows them to learn complex relationships between language and images. The applications of VLMs are extensive and span across various domains, including image captioning [7], [8], visual question answering [9], [10] and image generation [11], [12]. Their abilities also make them valuable tools in areas such as accessibility technology, autonomous systems, and digital content analysis.

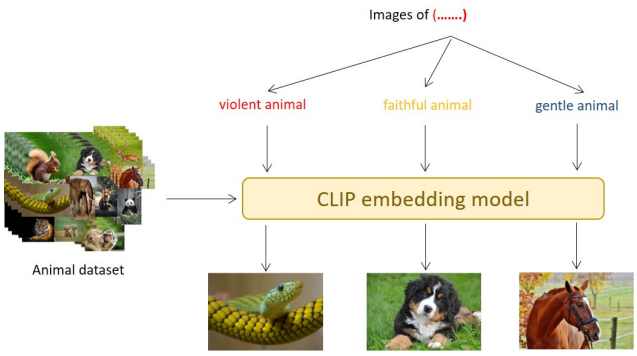


Fig. 1. Identifying animal biases in CLIP

However, the deployment of VLMs has raised concerns regarding the presence of biases embedded within them. Since they are trained on real-world data (often imbalanced or stereotypical), they inadvertently learn and propagate biases related to gender, race, and other protected attributes [13]. The primary types of biases identified in VLMs include gender bias, racial bias, and cultural bias [14].

The biases can result in significant social harms, including allocational harm, where opportunities and resources are unfairly distributed, and representational harm, where certain social groups are misrepresented or overlooked entirely [15]. Addressing the biases is crucial to ensure fair and ethical AI applications. The impact of biased LLM outputs on decision making processes in critical domains such as hiring, medical diagnosis, and criminal justice has been widely documented [16]. Understanding these biases is crucial for developing strategies to ensure equitable AI applications across various societal sectors. Various international frameworks and ethical AI guidelines emphasize fairness as a core criterion and underscore the necessity of bias mitigation to prevent discriminatory outcomes in AI applications [17].

Recently, more researchers are focusing their attention towards bias in VLMs [14], [18]–[20]. Despite extensive work on bias detection and mitigation, studies specifically addressing animal-related stereotypes in LLMs are sparse. However, addressing animal bias and stereotypes in cultural narratives is crucial mainly for two reasons - a)it impacts human perceptions, and b) it affects treatment of non-human species. Additionally, such narratives shapes the way the animals are being treated. The biases are often reinforced on the species that human categories superior to others based on human-centric values [21]. More importantly, such biases influence policies, ethical considerations, and AI-based decision making system [22]. Therefore, it is imperative to critically analyze and mitigate them to ensure ethical and fair representations of animals in digital and cultural spaces [23].

To that end, the proposed study extends the work on bias in VLMs by introducing an investigation into 'Animal Bias'. Animals often carry symbolic meanings in cultures; for instance, owls are wise or donkeys are foolish (see Fig. 1). Our approach involves curating a diverse image dataset of animals and evaluating the response of CLIP to prompts involving animals with culturally sensitive attributes. For that, we also introduce a new framework, called the Animal Bias Taxonomy (ABT), which categorizes various cultural stereotypes often associated with animals. Moreover, the study has used several performance metrics to assess the level of bias associated with animals. Since biased representations of animals can perpetuate misinformation or reinforce negative stereotypes, the current study is carried out with the goal that ethical considerations in AI should go beyond human-centered biases and should include cultural sensitivities around non-human entities such as animals.

The major contributions of the current work are as follows:

- We propose a novel taxonomy, ABT, that categorizes cultural bias related to animals, enabling a more comprehensive evaluation of animal bias in VLMs.
- An animal dataset is curated from different existing data sets and cleaned using the CLIP model prior to bias identification.
- Using ABT and animal dataset, we audit CLIP model to identify significant biases toward animals.
- Developed new metrics based on human experts' opinions.

II. Related works

Bias in large language models (LLMs) has been a growing concern within the field of NLP [24]–[26]. Bias can be induced in different ways. For example, gender bias manifests in the form of reinforcing traditional gender roles, such as associating women with domestic tasks and men with leadership roles, however, racial bias can lead to discriminatory misclassification. Additionally, cultural bias arises when VLMs fail to represent diverse cultural contexts fairly, often favoring Western-centric depictions and neglecting underrepresented regions and communities.

The literature suggests that researchers have predominantly focus on biases related to race, age, nationality, ethnicity, religion, political, sexual orientation, gender, occupation, and lifestyle [27], [28]. For instance, Cao and Bandara [29] investigated the stereotypical biases in proprietary and open-source LLMs. They carried out a comparison between GPT-4, Gemini-Pro, and LLaMA. They have investigated whether open-source models exhibited higher levels of stereotype scores compared to proprietary counterparts? They have found this true. Possibly because of differences in regulatory oversight and human reinforcement learning feedback. These insights are valuable for understanding how biases propagate in LLMs trained on diverse corpora.

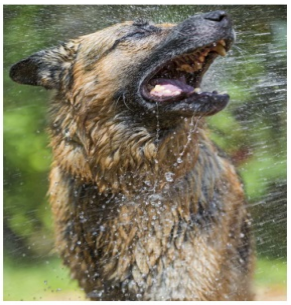
In addition to this, several studies have focused on quantifying biases in LLMs through benchmarking datasets. GPTEval [30] is a framework for automatically evaluating the performance of LLMs in terms of linguistic, reasoning, knowledge, and ethical tasks. Another extensive evaluation framework has been proposed in the NeurIPS benchmark dataset. This framework is meant for identifying fairness issues for various domains. In the same way, to test the model biases, authors have used StereoSet dataset [31]. Stereotype score and context association tests have been used for the purpose. Their primary focus has been biases related to race and religion. It is worth to mention that the majority of the studies have focused on unimodal text-based models [32]–[34]. Nadeem et. al [32] presented a large-scale dataset, StereoSet, designed to evaluate stereotypical biases in pretrained language models across four domains: gender, profession, race, and religion. They also introduced a novel evaluation framework, the Context Association Test (CAT), which includes intrasentence and intersentence tests to measure the presence of bias while assessing the models' language modeling capabilities. Abid et al. [33] explored the presence of bias in GPT-3, which consistently associates Muslims with violence at a significantly higher rate compared to other religious groups. Demidova et al. [34] investigated biases in GPT-3.5 and Gemini, across different languages and contexts using debate-based prompts. Their study focused on cultural, political, racial, religious, and gender biases by analyzing model responses to scenarios in Arabic, English, and Russian. A comprehensive survey of bias in large language models can be found in [35].

In addition to this, a few works have been proposed to mitigate bias in LLMs. For example, authors have suggested soft-prompt tuning [36] which reduces bias by adjusting input representation. Additionally, another work suggested in [37] uses causal mediation analysis and helps reducing bias by controlling for unintended correlations. Butter [38] have explored biases related to occupation in AI models. Authors have identified a systemic discrimination embedded within model outputs. Similarly, Ferrara [39] have explored fairness in AI models. They have investigated how a discrimination embedded within the model could induce bias for any domain. The findings of these studies are are further boosted by the outcome of [40] that suggests that addressing bias in multilingual and multicultural contexts are more complex.

(a) A church image under 'Turtle' class

(b) A lion with humans

(c) A crow image with overlay text



(d) A dog showing extreme emotion

Fig. 2. A few sample images that can impact the responses of CLIP model

Recently, more researchers are focusing their attention towards bias in VLMs [14]. Hamidieh et al. [41] analyzed the presence of social biases in CLIP by introducing a comprehensive taxonomy of biases called So-B-IT. The study investigated how CLIP associated harmful stereotypes with demographic groups, such as linking Middle Eastern men to terrorism. Cho et al. [42] investigated gender and skin tone biases in DALL-E generated images and highlighted the gender-based and racial bias in them. Similar to [32], Zhou et al. [43] introduced VLStereoSet, a dataset designed to measure social biases in VLMs by extending the text-based StereoSet dataset to the multimodal domain.

However, as stated in the previous section I, the biases in the LLMs towards animal have been underexplored [44]. Unbiased LLM towards animal is important to avoid justification of exploitation of animals as the historical framework of speciesism—where some animals are revered (e.g., lions symbolizing strength) while others are vilified (e.g., snakes as deceptive)—not only distorts human understanding of animal behavior but also justifies exploitation and discrimination against certain species [45]. This work aims to bridge this gap by investigating how LLMs perceive and generate content related to animals, potentially reinforcing cultural and societal stereotypes. However, addressing bias in LLMs presents several challenges, including the selection of unbiased training data and developing robust evaluation metrics.

Additionally, ensuring fairness across diverse linguistic and cultural contexts remain critical too [46], [47]. To that end, we have curated a dataset from CLIP, implying rigorous cleaning process to adequately assess the biases in large models. Further, an in-depth analysis of biases, their types and possible solutions have been discussed. In the subsequent section, we have detailed the dataset and their attributes.

III. Dataset curation and cleaning

For the current study, an image dataset was required which contain commonly stereotyped animals such as dogs, owls etc. We looked into existing literature but could not find images of all animals of interest for our research at one place. Therefore, we curated our own animal image dataset.

A. Animal dataset

We collected animal images from existing animal datasets [48] and publicly available sources (a complete list of sources is given in Appendix (see Table III)). The curated dataset comprised 21 distinct classes of animals with a total of 21,794 images. The classes included a diverse range of animals such as chameleons, frogs, goats, rabbits, monkeys, elephants, dogs, cats, lions, and snakes, among others (see sample images in Fig. 6). Each class was represented by a significant number of images, with the largest class (chameleons) containing 1,223 images, while the smallest class (donkey and wolf) had 1,000 images. The relatively balanced distribution ensures that no single class dominates the dataset.[1]

B. Dataset cleaning

While processing the images, it was realized that many classes contained images which are not relevant to our study (such as the image folder 'Turtle' contained an image of a church, see Fig. 2). Such images affected the ability of CLIP model to assign attributes to animals (the church image got the highest similarity score for the attribute 'Faithful'). Therefore, the dataset went through a cleaning process to make sure all the images were relevant to our study [49].

First, the images were checked to ensure they were valid and not damaged. To clean the dataset further, we used the CLIP model and supplied it a list of prompts to identify and remove images that could introduce noise in the results. To be concrete, we removed images based on the following criteria:

1) **Images containing objects other than animals:** Images with objects like cars, bulidings, furniture, or landscapes and which do not contain any animal were excluded. Such images might confuse the model during analysis by associating non-animal objects with certain animal categories.
2) **Images containing humans:** Sometimes we found images that included humans alongside animals. They were removed to avoid the influence of human presence on the results as we identified that images with humans were scored higher for attributes such as 'faithful animal'.

[1] https://kaggle.com/datasets/anas123siddiqui/animals

3) **Images containing text-overlays:** Text overlays on images, such as captions or watermarks, were removed because they could interfere with visual feature extraction and affect decision making.
4) **Animals displaying high emotions:** Images of animals showing extreme emotions (e.g., aggression, sadness, or joy) or fighting were excluded to prevent the VLM from being biased toward emotional portrayals.

To carry out the cleaning process, the CLIP model was supplied with prompts according to above-mentioned criteria and removed the top 200 images on the basis of similarity score. The distribution of the updated dataset is given in Table I. After cleaning, the animals in the dataset were mostly presented in a neutral manner.

IV. Methodology

The adopted methodology for the current work is described in this section and shown in Fig. 3.

A. CLIP

We used CLIP (Contrastive Language-Image Pre-training) model, developed by OpenAI [50]. It is a VLM designed to learn transferable visual representations by leveraging natural language supervision. Instead of training on pre-defined object categories, CLIP learns from a large-scale dataset of 400 million image-text pairs collected from the internet. The model uses a dual-encoder architecture, consisting of an image encoder and a text encoder, both of which are trained jointly using a contrastive learning objective. The training objective aligns image and text embeddings in a shared multimodal space, maximizing the similarity of correct image-text pairs while minimizing it for incorrect ones. Its design enables CLIP to generalize well to a wide range of downstream tasks without requiring fine-tuning. For the current work, we have used CLIP with a ViT-B/32 transformer architecture.

TABLE I
THE NUMBER OF IMAGES IN EACH CLASS BEFORE AND AFTER CLEANING PROCESS.

S. No	Class	Number of instances	
		Original dataset	Cleaned dataset
1	Bear	1008	896
2	Cat	1001	957
3	Chameleon	1223	1159
4	Crow	1018	919
5	Deer	1008	993
6	Dog	1032	949
7	Donkey	1000	964
8	Elephant	1038	1038
9	Fox	1013	971
10	Frog	1104	1099
11	Goat	1092	1031
12	Horses	1009	957
13	Lion	1019	971
14	Monkey	1054	1013
15	Mouse	1032	999
16	Owl	1001	919
17	Rabbit	1088	1082
18	Snake	1020	943
19	Squirrel	1007	996
20	Turtle_Tortoise	1027	988
21	Wolf	1000	975

B. Animal bias taxonomy (ABT)

To study animal bias comprehensively, we propose a Animal bias taxonomy (ABT) that categorizes attributes into three distinct yet interconnected categories: Behavior-Based, Intelligence-Based, and Physical Strength-Based (see Fig. 4). The taxonomy aims to provide a structured framework for analyzing the perception of animals based on stereotypical attributes often associated with them:

1) **Behavior-based:**The category includes attributes such as gentle, violent, and faithful to capture biases related to behavioral tendencies of animals. For instance, animals like horses and deers are stereotypically associated with gentleness, while wolves or lions may be perceived as violent. Similarly, dogs are commonly regarded as faithful due to their domestication history and bond with humans.
2) **Intelligence-based:**Attributes such as wise, foolish, clever, and dumb are put in this category. Perceptions of animal intelligence also often arise from in cultural narratives. For example, animals like owls are commonly associated with wisdom due to their symbolic representation in mythology, while donkey are often unfairly labeled as foolish or unintelligent.
3) **Physical strength-based:**The category includes two atributes: hardworking and strong, which capture biases tied to an animal's perceived physical capabilities. For example, donkeys and cows are labeled as hardworking due to their historical role in agriculture, while animals like horses or lions are revered for their strength.

While the proposed taxonomy is not exhaustive, it provides a foundational framework for categorizing and analyzing biases toward animals. Each category addresses a unique dimension of stereotyed perception. Moreover, the taxonomy is designed to be adaptable for future studies, such as adding subcategories for specific types of animals or regional variations in bias.

C. Prompt used

To evaluate bias, we designed a generic prompt as: "an image of a/an [attribute] animal". For each prompt, we used CLIP model to calculate the similarity between the text embedding of the prompt and the image embedding of a predefined dataset of animal images. The similarity score indicated the degree to which the attribute is associated with specific animal images in the dataset. By comparing similarity scores across prompts for different attributes, we can analyze patterns of biases present in CLIP.

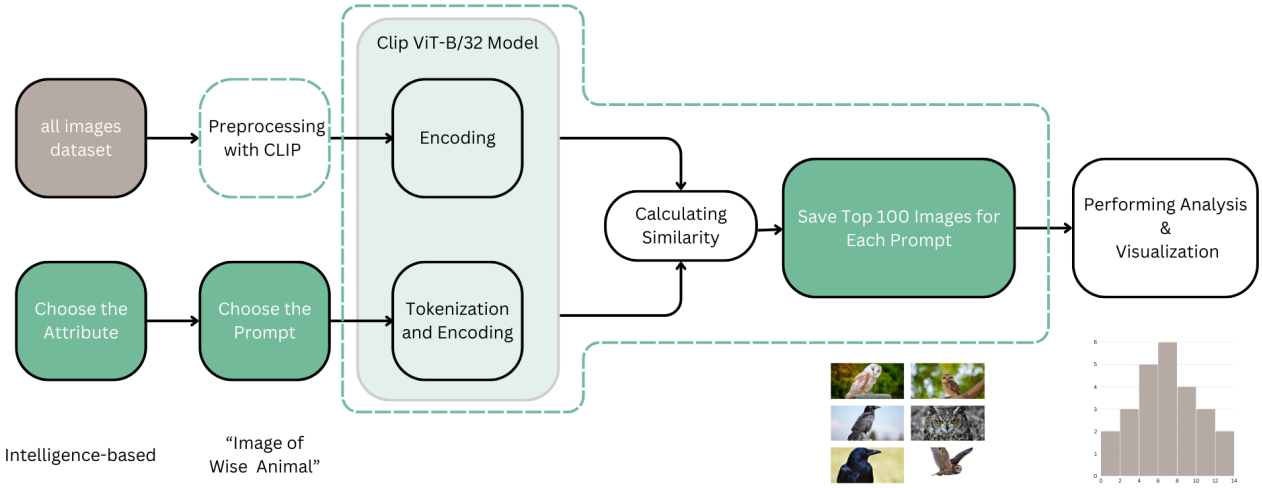


Fig. 3. Steps involved in the methodology adopted

D. Bias identification

For each category of animal bias in our taxonomy, we analyze the association of captions with animal images by retrieving the top-k samples with the highest similarity scores for each prompt. In our experiments, we use k=100. If the distribution of animal types is uniform across the top-k retrieved images, we infer that the CLIP model does not exhibit bias for the specific attribute. Conversely, if certain animals are significantly overrepresented or underrepresented in the top-k samples, we infer the presence of bias associated with that attribute. This process is repeated for all prompts, and the results were analyzed to identify recurring patterns.

E. Performance measures

Apart from frequency count, we have employed novel metrics to assess different facets of animal bias. Their details are as follows:

- **Expert-based evaluation:** In this approach, we consulted experts in literature and cultural studies, presenting them with the developed taxonomy. They were requested to identify common animals that are stereotypically associated with each attribute. Once received, we compared the experts' opinion with the responses of CLIP model.
- **Expert agreement precision (EAP):** EAP measures the degree of alignment between CLIP's top k predicted animals align with the k animals identified by experts for a given attribute. It is essentially a precision like score (precision@k) that evaluates the proportion of expert-defined stereotyped animals present in CLIP's top k predictions:

$$EAP@k = \frac{|E \cap C_k|}{k} \tag{1}$$

where E is a set (size = k) of expert-defined stereotyped animals for a given attribute and C represents the set of CLIP's top k predicted animals.

- **Expert Presence Score (EPS):** The EPS metric measures how many of the expert-defined stereotyped animals are present among the animals that have received the top m similarity scores in CLIP's predictions. It is similar to recall@m, but specifically evaluates how well expert-defined stereotypes are covered in CLIP's results:

$$EPS@m = \frac{|E \cap C_m|}{|E|} \tag{2}$$

For the current sudy, the experts were specifically instructed to provide exactly three animals (i.e. k=3 for EAP) per attribute to maintain consistency and prevent dilution of the categorization. For the EPS metric, we considered animals that appeared in CLIP's top 10 ranked results (i.e m=10 for EPS).

V. Findings

The responses of CLIP model were analyzed and the insights are shared next. Moreover, the distribution of frequencies of animals for each attribute can be found in Fig. 5.

A. Behavior-based biases

The horse is overwhelmingly associated with the gentle attribute (75% of the top 100 images). It aligns with cultural depictions in literature and media, where horses are often portrayed as majestic, calm, and cooperative animals. However, other naturally gentle animals, such as deer and rabbits had minimal representation. Goats (15%) and Donkeys (4%) also appear in this category, even though they are not often depicted as gentle. The highest-ranked animal in the violent category is the snake (25%), reinforcing its longstanding association with danger and deception in mythology, religion, and storytelling. The presence of wolves (14%) suggests that CLIP aligns with stereotypes of wolves as aggressive predators.

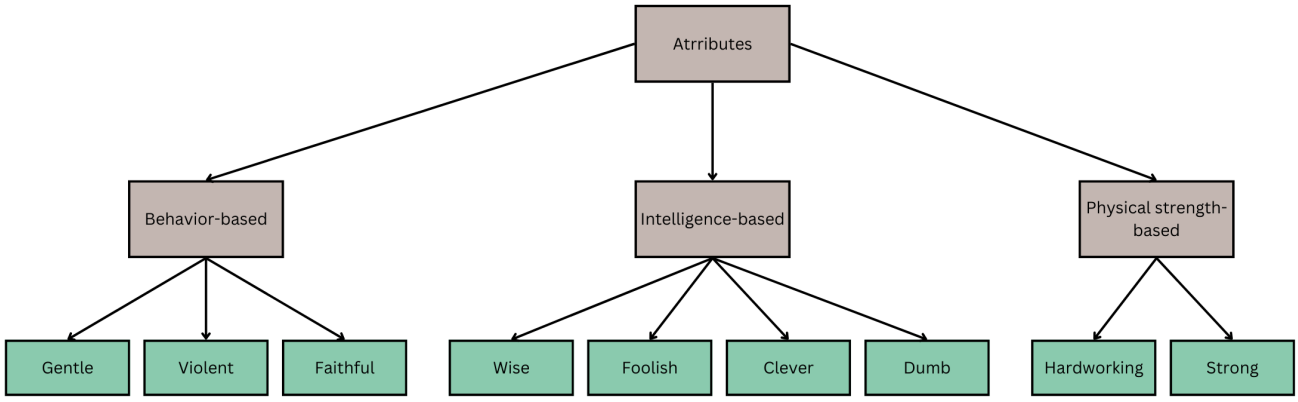


Fig. 4. Animal Bias Taxonomy

Mice (7%) and cats (5%) also appear the category, which is unexpected. Lions, typically seen as symbols of raw power and violence, are underrepresented (4%). Unsurprisingly, dogs dominate the "faithful" category with 79% representation which strongly reinforces their cultural association with loyalty and companionship. Goats (13%) are ranked second. Their presence might be due to dataset bias, where images of goats in close association with humans could have influenced CLIP's perception.

B. Intelligence-based biases

As expected, owls dominate the "wise" category (40%) which reflect their longstanding association with wisdom in across cultures, mythology, and literature. Squirrels are ranking second (35%) is unexpected. Their presence may be due to their problem-solving abilities in the wild, which could have been emphasized in training data. Though with lower representations, Monkeys (11%) and crows (7%) also appear in the list. The overwhelming presence of monkeys (58%) in the "foolish" category highlights a contradiction—monkeys are ranked high in both the wise and foolish categories, suggesting context-dependent bias. Donkeys (16%) being labeled as foolish directly reflects long-standing cultural stereotypes. CLIP shows association of chameleons (64%) with cleverness. It may stem from their adaptive camouflage ability, which might be interpreted as deceptive. Crows were ranked second (17%). Owls ranking lower (8%) is surprising, given their dominance in the "wise" category, suggesting that CLIP may percieve "cleverness" (practical intelligence) and "wisdom" (symbolic intelligence) differently. The donkey's overwhelming representation (60%) in the "dumb" category is a clear example of cultural bias. The stereotype has existed for centuries despite donkeys being highly trainable and intelligent animals. Deer (11%) and rabbits (7%) also appear in this category, likely due to their portrayal as simple-minded prey animals in storytelling.

C. Physical strength-based biases

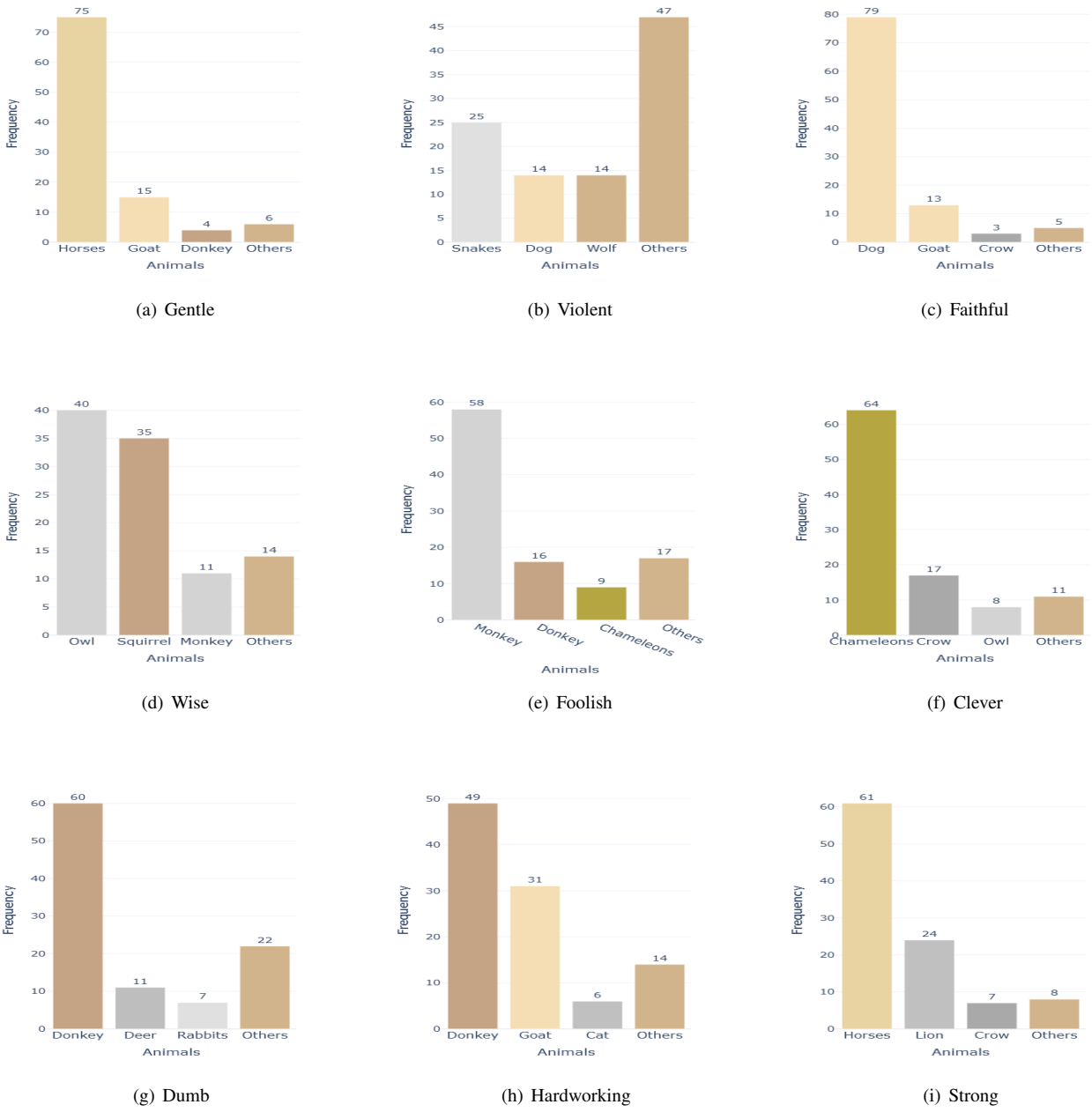
Donkeys (49%) and goats (31%) are strongly associated with being hardworking, which is expected given their historical use as work animals. Donkeys have been used in agriculture, transport, and labor-intensive tasks, reinforcing the stereotype. Goats' high representation could be due to their resilience in harsh environments, making them appear enduring and industrious. Horses (61%) ranking highest in strength is expected, given their historical role in transportation, agriculture, and warfare. The presence of lions (24%) also aligns with their reputation as powerful apex predators. However, crows (7%) appearing in this category is unusual. The low representation of wolves (2%) is surprising, considering their known physical endurance.

D. Discussion

Our analysis of CLIP's responses to animal-related attributes reveals a complex interplay between stereotypical associations and unexpected anomalies. While many of the results align with common cultural narratives about animals, there are also several instances where CLIP's predictions deviate from traditional stereotypes. We have discussed the key points next.

1) CLIP Reinforces Cultural Stereotypes: Across multiple attributes, CLIP's responses overwhelmingly reflect traditional stereotypes associated with animals. For example, the high association of owls with wisdom, dogs with faithfulness, donkeys with dumbness etc., aligns with their symbolic role in culture. Such findings indicate that CLIP has internalized widespread cultural biases and is reinforcing them in its outputs.

2) Anomalies: Despite common biases, CLIP also produced several unexpected associations such as squirrels ranking second in the "wise" category, crow at third place in "strong" category, cat as "hardworking" animals etc. It clearly indicates that CLIP does not rely solely on high-level semantic associations but also considers visual elements such as body posture,



(a) Gentle (b) Violent (c) Faithful

(d) Wise (e) Foolish (f) Clever

(g) Dumb (h) Hardworking (i) Strong

Fig. 5. The distribution of frequencies of animals among top-100 images

facial expressions, and environmental context when assigning attributes. To observe this trend further, we supplied CLIP a prompt with "weak" as the attribute. We found that CLIP assigned high similarity scores to images of frail or sickly-looking animals even when they belonged to species typically associated with strength, such as lions. It also suggests that despite the dataset cleaning process, there existed "noise" in the image corpus. It influenced CLIP's outputs which led to unexpected attributions.

3) Alignment with experts: Table II contains the results of metrics EAP and EPS which reveals varying degrees of alignment between CLIP's predictions and expert-defined stereotypes across different animal attributes. Strong alignment is observed in categories like faithful, clever, dumb, gentle, and strong, where CLIP's predictions closely match expert expectations. Moderate alignment is seen in attributes such as foolish and violent, where some expert-expected animals appear, but there are also unexpected associations like dogs and squirrels. Low alignment occurs in wise and hardworking, where the results of CLIP do not agree with expert opinions to a great extent.

TABLE II
RESULTS OF VARIOUS BIAS-RELATED METRICS

Category	Attribute	Experts list (k=3)	CLIP's top-3 Animals	EAP	Animals appearing in CLIP's top-10 scores	EPS
Behavior	**Wise**	horses, turtle, deer	horse, goat, donkey	0.34	horse, goat	0.34
	Foolish	snake, wolf, lion	snake, dog, wolf	0.67	wolf, bear, dog, mouse, deer, cat	0.34
	Gentle	dog, horse, goat	dog, goat, crow	0.67	dog, goat, crow	0.67
Intelligence	**Violent**	owl, turtle, crow	owl, squirrel, monkey	0.34	owl, crow, squirrle, monkey	0.67
	Faithful	donkey, frog, monkey	Monkey, donkey, chameleon	0.67	monkey, chameleon, donkey	0.67
	Clever	fox, monkey, chameleon	chameleon, crow, monkey	0.67	chameleon, crow, owl, frog, squirrle, mouse, monkey	0.67
	Dumb	donkey, deer, rabbit	donkey, deer, rabbit	1	donkey, horse, dog, chameleon, deer	0.67
Physical strength	**Hardworking**	horse, donkey, elephant	donkey, goat, cat	0.34	donkey, monkey, dog, cat	0.34
	Strong	elephant, lion, horse	horse,lion, crow	0.67	horse, lion, donkey, dog	0.67

VI. LIMITATIONS

We acknowledge several limitations. First, our analysis relies on a curated dataset of animal representations, which may not fully capture the breadth of their diversity. There were 22 classes of animals which can be increased for better results. Moreover, we considered various species of a specific animal as one class only (such as African and Indian elephants are significantly different but were put in the same class 'Elephant'). Additionally, our taxonomy of attributes is not exhaustive and may include other attributes that are often associated with animals.

We primarily focused on evaluating CLIP based on short textual prompts related to animals. However, real-world applications extend beyond simple text-image associations to more complex contexts, such as automated storytelling and educational tools. Future research should explore the impact of identified biases in more intricate scenarios. Another limitation is that our experiments are conducted using CLIP due to its prominence and accessibility in vision language research. While our findings provide valuable insights, they may not generalize across other models, such as DALL-E or BLIP, which employ different training data. Further investigations across a broader set of VLMs are necessary to validate the generalizability of our results.

Finally, our study focused on detecting and quantifying biases, but it does not assess their subjective impact. Conducting user studies and engaging with experts will be essential to gain deeper insights into how these biases affect perception and representation in AI-generated content. Despite the limitations, our work provides a foundational framework for future research aimed at mitigating cultural biases against animals in VLMs.

VII. CONCLUSION

This work highlights the presence of stereotypical biases in CLIP's vision language model when associating animals with culturally ingrained attributes. We have proposed an Animal Bias Taxonomy (ABT) and curated an animal dataset, by the virtue of these contributions, we systematically analyzed how CLIP responds to stereotype-laden prompts and found that its predictions frequently align with established cultural narratives.

This includes associating owls with wisdom, donkeys with foolishness, and dogs with faithfulness, among others. However, our findings also reveal that CLIP does not merely replicate these stereotypes but also exhibits unexpected associations, which suggests that its biases stem not only from linguistic conventions but also from image-based reasoning and dataset artifacts.

The observed anomalies, such as squirrels being categorized as wise, crows as strong, or cats as hardworking, indicate that CLIP integrates both textual attributes and visual characteristics in its decision making process. While our dataset underwent rigorous cleaning, residual biases within image distributions may have contributed to these unexpected attributions, emphasizing the need for careful dataset curation and interpretability in AI models. Our expert-aligned evaluation further underscores varying degrees of agreement between human-defined stereotypes and CLIP's outputs, reinforcing the necessity of auditing multimodal AI models for bias and inconsistencies.

Given the broader implications of such biases—ranging from misrepresentation of animals in AI-generated content to ethical considerations in conservation and education—this study underscores the importance of expanding fairness frameworks beyond human-centered biases. Future research should explore context-aware training methodologies, more diverse datasets, and bias-mitigation strategies to ensure that vision language models not only avoid reinforcing human stereotypes but also provide more accurate, unbiased representations of nonhuman entities.

ACKNOWLEDGMENT

This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005) and by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore. Mohammad Nadeem acknowledges the support provided by OpenAI through its Researcher Access Program (ID:0000006949). Amir Hussain acknowledges the support of the UK Engineering and Physical Sciences Research Council (EPSRC) Grants Ref. EP/T021063/1 (COG-MHEAR) and EP/T024917/1 (NATGEN).

REFERENCES

[1] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[2] M. Ge, R. Mao, and E. Cambria, "Discovering the cognitive bias of toxic language through metaphorical concept mappings," *Cognitive Computation*, vol. 17, 2025.

[3] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[4] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *CVPR*, 2021, pp. 3558–3568.

[5] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *NeurIPS*, vol. 35, pp. 25 278–25 294, 2022.

[6] R. Mao, Q. Liu, K. He, W. Li, and E. Cambria, "The biases of pre-trained language models: An empirical study on prompt-based sentiment analysis and emotion detection," *IEEE transactions on affective computing*, vol. 14, no. 3, pp. 1743–1753, 2022.

[7] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny, "Visualgpt: Data-efficient adaptation of pretrained language models for image captioning," in *CVPR*, 2022, pp. 18 030–18 040.

[8] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang, "Scaling up vision-language pre-training for image captioning," in *CVPR*, 2022, pp. 17 980–17 989.

[9] Z. Shao, Z. Yu, M. Wang, and J. Yu, "Prompting large language models with answer heuristics for knowledge-based visual question answering," in *CVPR*, 2023, pp. 14 974–14 983.

[10] J. Guo, J. Li, D. Li, A. M. H. Tiong, B. Li, D. Tao, and S. Hoi, "From images to textual prompts: Zero-shot visual question answering with frozen large language models," in *CVPR*, 2023, pp. 10 867–10 877.

[11] S. S. Sohail, F. Farhat, Y. Himeur, M. Nadeem, D. Ø. Madsen, Y. Singh, S. Atalla, and W. Mansoor, "Decoding ChatGPT: a taxonomy of existing research, current challenges, and possible future directions," *Journal of King Saud University-Computer and Information Sciences*, 2023.

[12] X. Li, X. Hou, and C. C. Loy, "When stylegan meets stable diffusion: a w+ adapter for personalized image generation," in *CVPR*, 2024, pp. 2187–2196.

[13] R. Wolfe and A. Caliskan, "Markedness in visual semantic ai," in *ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1269–1279.

[14] N. Lee, Y. Bang, H. Lovenia, S. Cahyawijaya, W. Dai, and P. Fung, "Survey of social bias in vision-language models," *arXiv preprint arXiv:2309.14381*, 2023.

[15] S. Barocas, K. Crawford, A. Shapiro, and H. Wallach, "The problem with bias: Allocative versus representational harms in machine learning," in *SIGCIS*. New York, NY, 2017, p. 1.

[16] J. Devlin *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," *NAACL*, 2019.

[17] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, "Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations," *Minds and machines*, vol. 28, pp. 689–707, 2018.

[18] M. Nadeem, S. S. Sohail, E. Cambria, B. W. Schuller, and A. Hussain, "Negation blindness in large language models: Unveiling the no syndrome in image generation," *arXiv preprint arXiv:2409.00105*, 2024.

[19] M. Nadeem, S. S. Sohail, L. Javed, F. Anwer, A. K. J. Saudagar, and K. Muhammad, "Vision-enabled large language and deep learning models for image-based emotion recognition," *Cognitive Computation*, 2024, accepted, in press.

[20] M. Nadeem, S. S. Sohail, E. Cambria, B. W. Schuller, and A. Hussain, "Gender bias in text-to-video generation models: A case study of Sora," *IEEE Intelligent Systems*, vol. 40, no. 3, 2025.

[21] A. Stibbe, *Ecolinguistics: Language, ecology and the stories we live by*. Routledge, 2015.

[22] T. Hagendorff, "Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods," *arXiv preprint arXiv:2303.13988*, vol. 1, 2023.

[23] P. Singer and Y. F. Tse, "AI ethics: The case for including animals," *AI and Ethics*, vol. 3, no. 2, pp. 539–551, 2023.

[24] E. Cambria, D. Rajagopal, D. Olsher, and D. Das, "Big social data analysis," in *Big Data Computing*, R. Akerkar, Ed. Chapman and Hall/CRC, 2013, ch. 13, pp. 401–414.

[25] L. Oneto, F. Bisio, E. Cambria, and D. Anguita, "Statistical learning theory and ELM for big social data analysis," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 45–55, 2016.

[26] Y. Li, Q. Pan, S. Wang, H. Peng, T. Yang, and E. Cambria, "Disentangled variational auto-encoder for semi-supervised learning," *Information Sciences*, vol. 482, pp. 73–85, 2019.

[27] A. Leidinger and R. Rogers, "How are LLMs mitigating stereotyping harms? learning from search engine studies," in *AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 839–854.

[28] ——, "Which stereotypes are moderated and under-moderated in search engine autocompletion?" in *ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 1049–1061.

[29] C. Cao and D. Bandara, "Evaluating stereotypical biases and implications for fairness in large language models," in *American Society for Engineering Education*, 2024.

[30] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTEval: A survey on assessments of ChatGPT and GPT-4," in *LREC-COLING*, 2024, pp. 7844–7866.

[31] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *ACL*, 2021.

[32] ——, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[33] A. Abid, M. Farooqi, and J. Zou, "Large language models associate muslims with violence," *Nature Machine Intelligence*, vol. 3, no. 6, pp. 461–463, 2021.

[34] A. Demidova, H. Atwany, N. Rabih, S. Sha'ban, and M. Abdul-Mageed, "John vs. ahmed: Debate-induced bias in multilingual llms," in *Arabic Natural Language Processing Conference*, 2024, pp. 193–209.

[35] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. K. Ahmed, "Bias and fairness in large language models: A survey," *Computational Linguistics*, 2024.

[36] J. Tian, D. B. Emerson *et al.*, "Soft-prompt tuning for large language models to evaluate bias," *arXiv preprint arXiv:2306.04735*, 2023.

[37] J. Vig *et al.*, "Investigating gender bias in language models using causal mediation analysis," in *NeurIPS*, 2020.

[38] S. Butter, "Unveiling gender bias in occupations," *Utrecht University*, 2024.

[39] E. Ferrara, "Should ChatGPT be biased? challenges and risks of bias in large language models," *arXiv preprint arXiv:2304.03738*, 2023.

[40] T. Naous *et al.*, "Measuring cultural bias in large language models," *arXiv preprint arXiv:2305.14456*, 2023.

[41] K. Hamidieh, H. Zhang, W. Gerych, T. Hartvigsen, and M. Ghassemi, "Identifying implicit social biases in vision-language models," in *AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 547–561.

[42] J. Cho, A. Zala, and M. Bansal, "Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models," in *International Conference on Computer Vision*, 2023, pp. 3043–3054.

[43] K. Zhou, Y. LAI, and J. Jiang, "Vlstereoset: A study of stereotypical bias in pre-trained vision-language models." Association for Computational Linguistics, 2022.

[44] T. Aman, M. Nadeem, S. S. Sohail, M. Anas, and E. Cambria, "Owls are wise and foxes are unfaithful: Uncovering animal stereotypes in vision-language models," *arXiv preprint arXiv:2501.12433*, 2025.

[45] O. Horta and F. Albersmeier, "Defining speciesism," *Philosophy Compass*, vol. 15, no. 11, pp. 1–9, 2020.

[46] L. Dixon *et al.*, "Measuring and mitigating unintended bias in text classification," in *AAAI*, 2018.

[47] K. Crawford *et al.*, "AI now 2019 report," *AI Now Institute*, 2019.

[48] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE TPAMI*, vol. 41, no. 9, pp. 2251–2265, 2018.

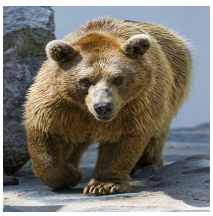
[49] D. Rajagopal, E. Cambria, D. Olsher, and K. Kwok, "A graph-based approach to commonsense concept extraction and semantic similarity detection," in *WWW*, 2013, pp. 565–570.

[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

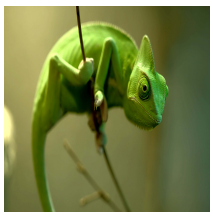
APPENDIX

TABLE III
SOURCES FROM WHERE ANIMAL BIAS DATASET IS CURATED.

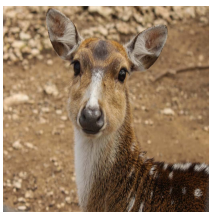
Sources	**Animals**
https://cvml.ista.ac.at/AwA2/ https://kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals https://kaggle.com/datasets/utkarshsaxenadn/animal-image-classification-dataset	Bear, cat, deer, dog, elephant, fox, horse, lion, rabbit, rat-mouse, monkey, squirrel
https://kaggle.com/datasets/harishvutukuri/dogs-vs-wolves	Wolves
https://kaggle.com/datasets/vencerlanz09/sea-animals-image-dataste?select=Turtle	Turtle-tortoise
https://kaggle.com/datasets/sameeharahman/preprocessed-snake-images	Snake
https://github.com/hohomsf/horse-donkey-classification/tree/master	Donkey
https://data.mendeley.com/datasets/4skwhnrscr/1	Goat
https://github.com/jonshamir/frog-dataset	Frog
Scrapping through publicly available sources	Owl, crow, chameleon


(a) Bear


(b) Cat

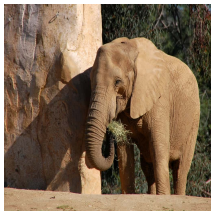

(c) Chameleon


(d) Crow


(e) Deer

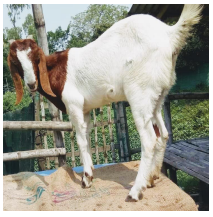

(f) Dog


(g) Donkey


(h) Elephant


(i) Fox


(j) Frog


(k) Goat


(l) Horse


(m) Lion


(n) Monkey

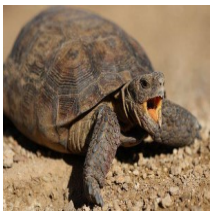

(o) Mouse


(p) Owl


(q) Rabbit


(r) Snake


(s) Squirrel


(t) Turtle


(u) Wolf

Fig. 6. A few sample images that can impact the responses of CLIP model