Are Vision Language Models Texture or Shape Biased and Can We Steer Them?

Paul Gavrikov[1] Jovita Lukasik[2] Steffen Jung[3,6] Robert Geirhos[4]
Bianca Lamm[1] Muhammad Jehanzeb Mirza[5] Margret Keuper[6,3] Janis Keuper[1,6]
[1] IMLA, Offenburg University [2] University of Siegen
[3] Max Planck Institute for Informatics, Saarland Informatics Campus
[4] Google DeepMind [5] ICG, Graz University of Technology [6] University of Mannheim

Abstract

Vision language models (VLMs) have drastically changed the computer vision model landscape in only a few years, opening an exciting array of new applications from zero-shot image classification, over to image captioning, and visual question answering. Unlike pure vision models, they offer an intuitive way to access visual content through language prompting. The wide applicability of such models encourages us to ask whether they also align with human vision — specifically, how far they adopt human-induced visual biases through multimodal fusion, or whether they simply inherit biases from pure vision models. One important visual bias is the texture vs. shape bias, or the dominance of local over global information. In this paper, we study this bias in a wide range of popular VLMs. Interestingly, we find that VLMs are often more shape-biased than their vision encoders, indicating that visual biases are modulated to some extent through text in multimodal models. If text does indeed influence visual biases, this suggests that we may be able to steer visual biases not just through visual input but also through language: a hypothesis that we confirm through extensive experiments. For instance, we are able to steer shape bias from as low as 49% to as high as 72% through prompting alone. For now, the strong human bias towards shape (96%) remains out of reach for all tested VLMs.

1. Introduction

As the old adage goes, all models are wrong, but some are useful. Similarly, all (machine learning) models are biased, but according to the no free lunch theorem, only some biases are useful [3]. While most biases observed in language models reflect social particularities that originate in unbalanced training data, vision models have been shown to be exceptionally misaligned with human perception, referred to as the *texture vs. shape bias* [1]. Specifically, when recognizing objects, an object's shape often plays a minor role to models while the texture is strongly preferred. Humans instead predominantly decide by the object's shape (96% shape over texture decisions). This finding has received significant attention [4–7], because it shows that our vision models, while approaching near-human levels in accuracy on specialized tasks, poorly approximate fundamentals of human perception.

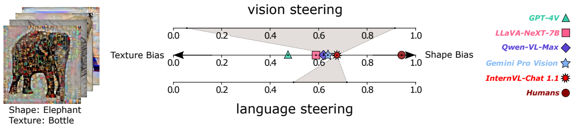

Figure 1. **Unlike many unimodal models, vision language models (VLMs) prefer shape over texture for object recognition, but not to the same extent as humans. Further, we find that the (visual) texture/shape bias [1] can be steered through language alone, albeit not to the extent as through vision.** Here we visualize the texture/shape bias of some exemplary VLMs, and highlight the steerability of InternVL-Chat 1.1 [2]. This paper does not show details on vision steering for space reasons. The reader can assume that the shape bias can be steered by image modification to almost either end - at impact on accuracy.

As the current generation of deep learning models is increasingly multimodal, it is often unclear whether these models inherit certain biases from their text or vision encoders, or whether those interact and are combined through multimodal fusion. Specifically for the (vision-only) texture vs. shape bias, we are interested in how this bias is influenced by language. If language is indeed able to influence a vision-only bias, this offers the possibility of aligning and generally steering biases simply and intuitively via prompts.

We study the texture vs. shape bias and its steerability in recent vision language models (VLMs). Here we use VLM to refer to text/vision-input models that output text generated by a large language model (LLM). Our investigation shows that the model's inherent bias towards texture is far less pronounced in VLMs than in most previously studied vision-only models. As shown in Fig. 1, most VLMs decide by shape more often than by texture, even when prompted in a neutral way (while not approaching the human shape bias). Further, we can confirm that VLMs learn to understand the visual concepts of *shape* and *texture*, such that they allow steering of the prediction to some extent by simple prompt modifications.

2. Methodology

A cornerstone of our analysis is the measurement of texture/shape bias in (LLM-based) VLMs when performing tasks that require some level of image classification. We measure the texture/shape bias of VLMs in two common tasks: *visual question answering (VQA)* [8] where we seek to obtain a zero-shot classification [9] of the object, and *image captioning* [10] where we look for an accurate but brief description of the image.

Measuring Texture/Shape Bias. We use the texture-shape cue-conflict classification problem (*cue-conflict*) [1] consisting of 1,280 samples with *conflicting* shape and texture cues synthetically generated via a style transfer model from ImageNet [11] samples (see Fig. 1 for examples). The shape and texture classes belong to 16 super-classes of ImageNet. We measure *accuracy*, by predictions that match the shape or texture label. We use the definition of *shape bias* [1], which is defined by the ratio of shape decisions over accurate decisions.

Image Captioning. In this task, we are instructing models to generate brief descriptions ("Describe the image. Keep your response short."). We specifically request the model to provide a short response to encourage it to single out the most crucial aspects of the image according to its judgment. Additionally, this has the benefit of faster inference. As the responses are open-ended we rely on zero-shot classifications of the generated description via embeddings of ember-v1 to marginalize out the most descriptive class. Note that captions may refer to multiple classes or be generic. Thus, we additionally prompt an LLM (Mixtral 8x7B Instruct v0.1) to provide a (possibly empty) list of all mentioned classes in the description.

VQA Classification. Following the questioning style in LLaVA [12], we ask the model "Which option best describes the image?" and provide an alphabetic enumeration of all class labels in the style "A. airplane". For a simpler response extraction, we end the prompt by instructing the model to answer with only the letter corresponding to the correct answer ("Answer with the option's letter from the given choices directly."). Compared to captioning, this is similar to the discrimination in ImageNet [11] image classifiers in the sense that it only allows the model to respond with a single class and does not provide an option to not answer - if models follow the instruction. We apply some simple post-processing to models that don't follow but avoid heavy post-processing and consider individual answers as wrong if they are not recoverable. In most cases, the ratio of these is negligible.

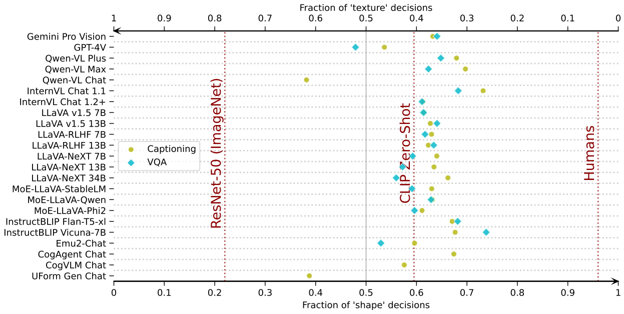

Figure 2. **Most VLMs are slightly shape-biased but some models show differences when asked to describe an image compared to VQA.** We measure the shape bias on the cue-conflict dataset [1]. For reference, we also provide measurements on ResNet-50 [13], zero-shot classification (CLIP ViT-L/14 [9]), and a human average (over 10 subjects [1]).

3. Are VLMs Biased towards Texture?

We start our experimental evaluation by measuring the shape bias in the VQA and captioning tasks, using a collection of diverse VLMs reflecting the multitude of research directions. These models include connections of common pretrained CLIP encoders and LLMs, mixture-of-expert-LLMs, optimized architectures for resource-constrained systems, finetuning with RLHF, massive vision encoders. Additionally, we survey commercial, closed-source models like Gemini Pro Vision 1.0, GPT-4V, and Qwen-VL Plus/Max where access is limited to APIs and few details are known.

The results in Fig. 2 paint a fairly uniform picture. Across different models and on two different tasks, most VLMs seem to perform relatively similarly in terms of shape bias. Shape bias of VLMs is still significantly lower than in humans (96%), but overall higher than in typical image-only discriminative classifiers (*e.g.*, 22% for an ImageNet-trained ResNet-50 [1, 13]). For most models, shape bias is higher than the shape bias of CLIP ViT-L/14 [9] (60%) which is a common vision encoder used in many of our tested models.

On average, we observe a slightly higher shape bias for the image captioning task at some cost in accuracy (mostly due to generic captions that do not mention any objects) and vice-versa. Shape bias can be significantly low (38.2%) for some outliers but generally ranges from 54.1 - 73.2%. Most captions do not refer to multiple classes (92.2% on average), but a significant ratio of captions is generic (31.9% on average). For VQA shape bias ranges from 52.9 - 73.8%. In this task, we had to remove some models as they did not follow instructions. Some of these models are on the lower end of the shape bias in captioning, which raises the question of whether there might exist a correlation between instruction tuning quality and shape bias (to answer this question, more samples would be needed for a conclusive answer).

Which models are the most shape-biased? The strongest shape bias is observed in InstructBLIP Vicuna-7B [14] for VQA, but the model generally shows a lower accuracy compared to other models. A more accurate model is InternVL-Chat 1.1 [2] which ranks second place for VQA but first for captioning.

GPT-4V is an outlier. Given that GPT-4V often achieves SOTA performance and is considered an important baseline, it has a surprisingly poor accuracy compared to most other models - mostly due to refusal to answer: 131/1280 VQA conversations, i.e., roughly 10% which is substantially higher than the refusal rate of all other models ($<$ 1%). Note that refusal rates do not affect the shape bias measurement. GPT-4V is also the model with the largest amount of generic image captions (60.4%). We acknowledge that other prompts may have led to other results, however, the result is noteworthy, as the other VLMs mostly behave well under the same prompts. Interestingly, GPT-4V is also significantly more texture-biased than most models in both tasks.

Does scale matter? LLM capacity does not seem to correlate with shape bias and unpredictably skews the shape bias by a few percent in each way as can be seen in Qwen-VL, LLaVA v1.5/NeXT/RLHF, or InternVL. Similarly, the overall largest models do not have the highest shape bias. However, scale *usually* improves accuracy.

Does RLHF align shape bias? RLHF-tuned VLMs are still rare at this point and we only have three samples. On both LLaVA-RLHF [15] models we see no changes in comparison to the default LLaVA models. GPT-4V [16] (though it is unclear if vision was also RLHF trained) shows one of the lowest shape biases in our study, but we do not know how the base model ranks. Overall it is hard to derive a conclusive answer, but it seems that at least RLHF does *not necessarily guarantee* an alignment of visual preferences.

4. Steering the Bias in Language

Our previous results suggest that VLMs learn a connected multimodal understanding of shape and texture because the shape bias differs from the vision encoder alone. This opens the question of whether visual biases can be influenced through text processing in these models. We test this hypothesis by recording texture/shape bias as a function steering them via text through prompt engineering.

Bias steering through hand-crafted prompting. We start by asking VLMs to specifically identify either the "shape" or the "texture" category in a given cue-conflict image. This *hand-crafted biased prompting* approach does indeed steer shape bias to a certain degree: As shown in Fig. 3, prompting can steer a visual bias (*without* significantly affecting accuracy). Neutral prompts perform often similarly to shape-biased prompts, whereas texture-biased prompts deviate more significantly. This suggests that models may be more inclined to use shape by default, but also have access to a certain amount of texture information which can be accessed through biased prompting.

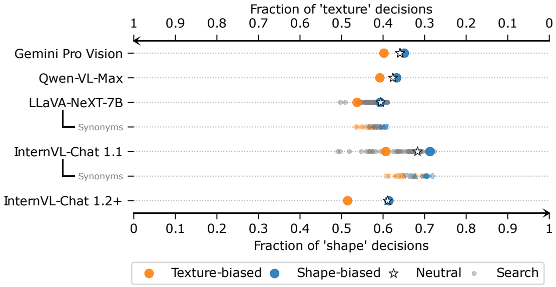

Figure 3. **Prompts can steer the texture/shape bias.** We test the same texture/shape-biased instructions on multiple models, showing that these can already shift some decisions (usually in favor of texture). For InternVL 1.1 and LLaVA-NeXT 7B we additionally test the understanding of texture/shape by using synonyms. Furthermore, we use an LLM to automatically search for specific prompts to optimize in either direction.

For two models (InternVL-Chat 1.1 [2] and LLaVA-NeXT 7B [17]) we additionally replace the terms texture/shape by strong synonyms obtained from *Thesaurus.com*. Synonyms of either term can steer shape bias as well to a certain degree and cluster well with their parents. We observe more variance for "texture" synonyms, as "texture" is overloaded by different meanings (*e.g.*, some synonyms like "feeling", "taste", or "touch" are unrelated to texture in vision). In contrast, "shape" is a fairly well-defined term.

While the effect of steering by language is systematically visible, language steering alone does not fundamentally change the reliance on the underlying cue. This effect does not appear to be a limitation of LLM capacity: We performed an additional study on InternVL-Chat 1.2+ (34B vs. 13B) but did not obtain evidence that larger LLMs guarantee more steerability. Interestingly, our findings of steering by language show parallels to a human psychophysical laboratory experiment. [1] conducted control experiments in which humans were either asked to identify the shape while ignoring texture, or conversely to identify the texture while ignoring the shape. This "human prompt steering" worked, but only to a certain extent: When humans were tasked to ignore the shape, the human shape bias decreased from 96% (neutral instruction) only to approx. 70% shape bias (texture-biased instruction). Our models behave somewhat similarly: their visual shape bias can be steered through prompting, but it appears hard for them to completely go against their default visual bias.

Bias steering through automated prompt engineering. Above we observed that hand-crafted prompts can steer visual biases, but only to a limited extent. Does this indicate a limit on how much language/prompting can influence biases, or merely reflect that the handcrafted prompts were chosen suboptimally? To investigate this question, we tested an *automatically crafted prompt*. This is achieved by employing an LLM as optimizer [18] to continuously generate new prompts in natural language targeting to maximize either shape or texture bias in a feedback loop. We provide the LLM feedback about the achieved accuracy and shape bias.

The results are shown in Fig. 3 in gray and denoted as "search". We observe that for both `LLaVA-NeXT-7B` and `InternVL-Chat 1.1`, automatically generated prompts exceed the manually crafted biased prompts in terms of their effectiveness to increase texture bias, and roughly match them when it comes to increasing shape bias. For `InternVL-Chat 1.1` the delta between both extremes is 23.3%, which can only serve as a lower bound and is likely improvable by a better design of the LLM task (or using other optimizers). In line with hand-crafted prompts, overall accuracy does not change considerably. We should also note that the optimization here is done with respect to the cue-conflict test set; this is simply done as a proof of concept to show that there are prompts that can substantially influence visual biases and not to claim a SOTA shape bias. It would be interesting to test whether the generated prompts generalize to other models or datasets.

Taken together, language-based steering clearly has a more reduced effect compared to visual steering. That said, unlike visual steering through image modifications which substantially hurts accuracy, our prompting strategy barely affects accuracy. Additionally, the method may be more intuitive for practitioners who can access or change the prompt as desired.

5. Conclusion

In this study, we investigated texture/shape bias in VLMs. Through the lens of this specific bias, we are able to assess whether visual biases are inherited from vision encoders or modified through text processing. We find that the latter is the case: VLMs are often more shape-biased than their vision-only backbones. Beyond that, an intriguing aspect of VLMs is that the task/response can be steered through not just vision but also through language. Our experiments have demonstrated that VLMs have learned a multimodal association between the terms "shape"/"texture", their synonyms, and their respective visual concepts to some extent. Taken together, we see that when it comes to multimodal models, the sum is greater than the parts, as a visual bias (texture/shape bias) can be steered to a certain degree through both manual and automated prompt engineering, alone. Similar to the active research line on prompt engineering for CLIP, it remains to be seen if prompts exist that can steer texture/shape bias to the same extent as vision can. While this aligns with human behavior in psychophysical studies [1], it remains unclear if this is a reflection of a learned bias from human-annotated data or a spurious correlation. We plan to explore this further in future studies.

Lastly, despite experiments on an extensive and diverse collection of VLMs, we have seen a (surprisingly) homogenous landscape in terms of texture/shape bias and steerability. We are curious to see how our insights apply under more radical training/architecture changes and to other visual biases. We encourage the community to provide more open-source implementations without which studies like ours would not be feasible.

References

[1] R. Geirhos *et al.*, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.," in *ICLR*, 2019.

[2] Z. Chen *et al.*, "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," 2024.

[3] D. Wolpert *et al.*, "No free lunch theorems for optimization," *IEEE TEVC*, 1997.

[4] K. Hermann *et al.*, "The Origins and Prevalence of Texture Bias in Convolutional Neural Networks ," in *NeurIPS*, 2020.

[5] M. A. Islam *et al.*, "Shape or Texture: Understanding Discriminative Features in CNNs," in *ICLR*, 2021.

[6] M. M. Naseer *et al.*, "Intriguing Properties of Vision Transformers," in *NeurIPS*, 2021.

[7] A. Subramanian *et al.*, "Spatial-frequency channels, shape bias, and adversarial robustness," *NeurIPS*, 2024.

[8] S. Antol *et al.*, "VQA: Visual Question Answering," in *ICCV*, Dec. 2015.

[9] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *ICML*, 2021.

[10] O. Vinyals *et al.*, "Show and Tell: A Neural Image Caption Generator," in *CVPR*, June 2015.

[11] J. Deng *et al.*, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[12] C. Li *et al.*, "Multimodal Foundation Models: From Specialists to General-Purpose Assistants," 2023.

[13] K. He *et al.*, "Deep Residual Learning for Image Recognition," 2015.

[14] W. Dai *et al.*, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," in *NeurIPS*, 2023.

[15] Z. Sun *et al.*, "Aligning Large Multimodal Models with Factually Augmented RLHF," 2023.

[16] OpenAI, "GPT-4 Technical Report," 2023.

[17] H. Liu *et al.*, "LLaVA-NeXT: Improved reasoning, OCR, and world knowledge," Jan. 2024.

[18] C. Yang *et al.*, "Large Language Models as Optimizers," in *ICLR*, 2024.