Contrast Sensitivity Function of Multimodal Vision-Language Models

Pablo Hernández-Cámara[a*], **Alexandra Gomez-Villa**[a], **Jose Manuel Jaén-Lorites**[b], **Jorge Vila-Tomás**[a]
Jesus Malo[a], **Valero Laparra**[a]
[a] Image Processing Lab, Universidad de Valencia, Paterna, Spain
[b] Center for Biomaterials and Tissue Engineering Universitat Politecnica de Valencia, Valencia, Spain
[*] Corresponding author: pablo.hernandez-camara@uv.es

Abstract

Assessing the alignment of multimodal vision-language models (VLMs) with human perception is essential to understand how they perceive low-level visual features. A key characteristic of human vision is the contrast sensitivity function (CSF), which describes sensitivity to spatial frequency at low-contrasts. Here, we introduce a novel behavioral psychophysics-inspired method to estimate the CSF of chat-based VLMs by directly prompting them to judge pattern visibility at different contrasts for each frequency. This methodology is closer to the real experiments in psychophysics than the previously reported. Using band-pass filtered noise images and a diverse set of prompts, we assess model responses across multiple architectures. We find that while some models approximate human-like CSF shape or magnitude, none fully replicate both. Notably, prompt phrasing has a large effect on the responses, raising concerns about prompt stability. Our results provide a new framework for probing visual sensitivity in multimodal models and reveal key gaps between their visual representations and human perception.

Keywords: Contrast Sensitivity Function; Vision-Language Models; Perceptual Alignment; Multimodal Perception; Computational Neuroscience; Psychophysics-inspired Evaluation

Introduction

Assessing the modern deep learning models' alignment with human vision and cognition is essential to understand their perceptual properties. One key measure in visual neuroscience is the contrast sensitivity function (CSF), which quantifies how sensitivity to visual patterns varies for different spatial frequencies (Campbell & Robson, 1968). While previous studies have measured the CSF of convolutional nets (Q. Li et al., 2022; Akbarinia, Morgenstern, & Gegenfurtner, 2023) and transformer-based vision models (Akbarinia et al., 2023; Cai et al., 2025), no work has yet explored chat-vision-language models (CVLMs)—multimodal systems capable of integrating vision and language through generative responses.

Traditional approaches to measuring CSF in deep networks rely on explicit readouts from internal representations. Some methods use classification-based readouts (Akbarinia et al., 2023), where a classifier is trained to detect gratings at different contrasts, conflating model perception with the classifier's performance. Others rely on Euclidean distance (Q. Li et al., 2022) or cosine similarity (Cai et al., 2025) in feature space, assuming that contrast perception corresponds to embedding differences, which imposes a potentially unnatural metric.

Here, we propose a behavioral approach that mirrors human psychophysical experiments. Instead of relying on internal representations, we directly query the model about visual stimuli. Specifically, we present CVLMs with structured images at different contrasts for each spatial frequency, asking whether they appear "flat" or contain a "pattern." By systematically varying the prompts and analyzing response consistency, we estimate the model's CSF in a way that is both naturalistic and interpretable. See figure 1 for a sketch of the method. We validate our approach on small open-source models ($\leq$ 7B parameters) and plan to apply it to larger available models.

Methods

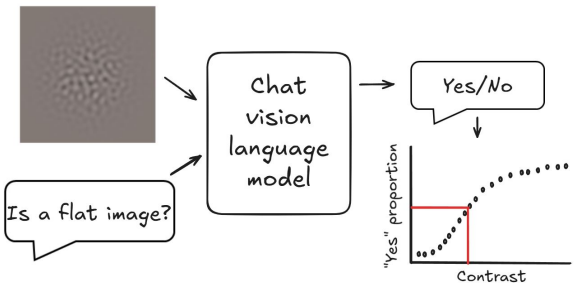

Figure 1: Method overview: Images and questions are used as input to the chat-vision-language models. The "Yes/No" answer proportion is computed as a function of the image contrast for each frequency.

Stimuli: To estimate the CSF of CVLMs, we generated 256×256 images by applying bandpass filters to white noise in the achromatic channel of ATD color space, using the Fourier domain. This yields perceptually equivalent but statistically distinct stimuli, improving robustness over sinusoidal gratings. Images were normalized to a fixed mean luminance and converted to RGB for model compatibility. Each image spans 4×4 visual degrees, aligning with the extent of human foveal vision.

Task Design: Following human psychophysics, each model was presented with an image and a prompt (as "*<image> Is there a pattern on the image?*") to classify it as containing a "pattern". To check prompt robustness, we systematically varied the prompts with 10 synonyms of "pattern", 10 synonyms of the adjective "visible", and 5 prompts varying the words order. Each contrast-frequency-prompt condition was tested 10 times to generate psychometric functions.

CSF Estimation: For each spatial frequency, we fit a lo-

gistic function to the proportion of affirmative responses as a function of contrast (the so-called psychometric functions), meaning that the model sees a pattern in the image. Then, we estimated the threshold contrast at which the model correctly detects the pattern 50% of the times. The CSF was computed as 1/threshold contrast, yielding a sensitivity curve directly comparable to the human CSF.

Models Tested: We evaluated multiple open-source CVLMs with equal or fewer than 7B parameters: Llava1.5-7B (Liu et al., 2023), Blip2-7B (J. Li et al., 2023), InstructBlip-Vicuna-7B (Liu et al., 2023), and Qwen2.5Vl-3B (Bai et al., 2025). Each model received identical images and prompts. Future work will extend this approach to larger models (ChatGPT, Gemini-Pro) and to contrastive-trained models (CLIP, SigLIP) to explore differences in visual sensitivity across architectures, training goals and model sizes.

Results

Figure 2 compares the contrast sensitivity functions (CSFs) of the analyzed chat-vision-language models, averaged over 25 different prompts, to the human CSF. The human CSF exhibits a characteristic bandpass shape, peaking around 4–6 cycles per degree (cpd) before declining at higher frequencies. LLaVA-1.5-7B is the model that most closely resembles human CSF in overall sensitivity, while Qwen2.5VL-3B, despite lower sensitivity, is the one that better captures the human CSF shape. Blip2-7B and InstructBlip-Vicuna-7B obtained much flatter CSFs, lacking clear peaks. In general, most models showed higher sensitivity than humans at high spatial frequencies, indicating a focus on fine details.

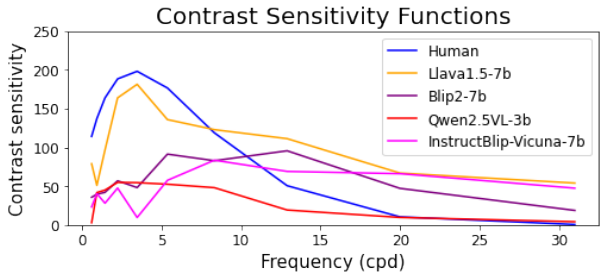

Figure 2: **CSFs results:** Human and Chat-VL-model contrast sensitivities averaged over the 25 prompts.

The qualitative observations are confirmed by the metrics in Table 1. Qwen2.5VL-3B has the highest shape Pearson correlation with the human CSF, while LLaVA-1.5-7B has the lowest RMSE, indicating close absolute values. Blip2-7B and InstructBlip-Vicuna-7B performed poorly in both shape and magnitude alignment.

Model	$P_\rho \uparrow$	RMSE $\downarrow$
Llava1.5-7B	0.70	50.4
Blip2-7B	0.20	90.8
InstructBlip-Vicuna-7B	-0.46	105.8
Qwen2.5VL-3B	0.85	97.6

Table 1: **CSF similarity metrics across models:** Pearson correlation and RMSE between the human CSF and the different Chat-VL-model CSFs averaged over the 25 prompts.

The earlier results averaged model CSFs across 25 prompts, but this overlooks prompt variability, which can significantly affect responses. A more accurate analysis computes correlation and RMSE for each prompt-specific CSF, then reports the mean and standard deviation to assess both human CSF alignment and response stability.

Table 2 shows these results. Regarding correlation, Blip2-7B and InstructBlip-Vicuna-7B have weak positive mean correlations, suggesting limited alignment with human trends. LLaVA-1.5-7B and Qwen2.5VL-3B show negative correlations, indicating deviation from human CSF, though they are more consistent across prompts, as reflected by lower standard deviations. Blip2-7B, by contrast, shows high variability, suggesting strong prompt dependence. RMSE results follow a similar pattern. InstructBlip-Vicuna-7B and Qwen2.5VL-3B have the lowest RMSEs, indicating values closer to human perception. LLaVA-1.5-7B has the highest RMSE, while Blip2-7B again shows the most variability. Overall, while some models may resemble human CSFs in certain cases, their strong dependence on prompt phrasing raises concerns about consistency and robustness.

Model	$P_{\rho,mean}$	$P_{\rho,std}$	$RMSE_{mean}$	$RMSE_{std}$
Llava1.5-7B	-0.24	0.16	117.3	98.8
Blip2-7B	0.22	0.40	68.9	111.5
InstructBlip-7B	0.22	0.32	59.7	77.4
Qwen2.5VL-3B	-0.24	0.24	45.7	93.8

Table 2: **CSF similarity metrics across models and prompt sets:** Mean and standard deviation (std) of the Pearson correlation and RMSE between each model's CSF and the human CSF across 25 distinct prompts.

Conclusions

In this work we introduce a novel approach for evaluating contrast sensitivity in chat-based vision-language models using a psychophysics-inspired framework. Our findings reveal that while models demonstrate varying degrees of sensitivity to visual contrast, none fully replicate the shape or stability of the human contrast sensitivity function (CSF).

Between the analyzed models, LLaVA-1.5-7B exhibits high overall sensitivity but shows significant prompt dependency, leading to inconsistent estimates. Qwen2.5VL-3B, although less sensitive in absolute terms, displays more stable CSF trends across different prompts. Blip2-7B and InstructBlip-Vicuna-7B, on the other hand, struggle with both shape alignment and stability, highlighting limitations in how these models process fundamental perceptual information. Prompt variability plays a crucial role in model performance, especially for models like Blip2-7B whose responses fluctuate widely with minor changes in phrasing. This sensitivity raises concerns about their internal consistency and reliability for tasks requiring fine-grained visual understanding.

Overall, this work offers a foundation for systematically probing low-level visual properties in multimodal models. It underscores the importance of developing prompt-invariant evaluation techniques and encourages further research into

aligning model perception with human vision—not only in task performance but also in fundamental perceptual processes.

Acknowledgments

This work was supported in part by MCIN/AEI/FEDER/UE under Grants PID2020-118071GB-I00 and PID2023-152133NB-I00, by Spanish MIU under Grant FPU21/02256 and in part by Generalitat Valenciana under Projects GV/2021/074, CIPROM/2021/056, and by the grant BBVA Foundations of Science program: Maths, Stats, Comp. Sci. and AI (VIS4NN). Some computer resources were provided by Artemisa, funded by the EU ERDF through the Instituto de Física Corpuscular, IFIC (CSIC-UV).

References

Akbarinia, A., Morgenstern, Y., & Gegenfurtner, K. R. (2023). Contrast sensitivity function in deep networks. *Neural Networks*, *164*, 228–244.

Bai, S., et al. (2025). Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Cai, Y., et al. (2025). Do computer vision foundation models learn the low-level characteristics of the human visual system? *arXiv preprint arXiv:2502.20256*.

Campbell, F. W., & Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of physiology*, *197*(3), 551.

Li, J., et al. (2023). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730–19742).

Li, Q., et al. (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision*, *22*(6), 8–8.

Liu, H., et al. (2023). Visual instruction tuning. *Advances in neural information processing systems*, *36*, 34892–34916.