多模态视觉-语言模型的对比敏感度函数

ablo Hernández-Cámara[a*], Alexandra Gomez-Villa[a], Jose Manuel Jaén-Lorites[b], Jorge Vila-Tomás[a], Jesus Malo[a], Valero Laparra
[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

摘要

评估多模态视觉语言模型（VLMs）与人类感知的对齐程度对于理解它们如何感知低级视觉特征至关重要。人类视觉的一个关键特征是对比度敏感度函数（CSF），它描述了在低对比度下对空间频率的敏感度。在此，我们引入了一种新颖的行为心理物理学启发的方法，通过直接提示聊天式 VLMs 来判断不同频率下的图案可见性，以估计其 CSF。与之前报道的方法相比，这种方法更接近心理物理学的真实实验。我们使用带通滤波的噪声图像和多样化的提示集，评估多种架构的模型响应。我们发现，尽管一些模型近似人类的 CSF 形状或幅度，但没有一个能够完全复制两者。值得注意的是，提示的措辞对响应有很大影响，这引起了对提示稳定性的担忧。我们的结果为探索多模态模型的视觉敏感性提供了新的框架，并揭示了它们的视觉表示与人类感知之间的关键差距。

关键词: 对比敏感度函数; 视觉语言模型; 感知对齐; 多模态感知; 计算神经科学; 心理物理学启发的评估

介绍

评估现代深度学习模型与人类视觉和认知的匹配情况对于理解其感知特性至关重要。视觉神经科学中的一个重要指标是对比敏感度函数（CSF），该函数量化了对视觉模式的敏感度如何随着不同空间频率 (Campbell & Robson, 1968) 的变化而变化。虽然之前的研究已经测量了卷积网络 (Q. Li et al., 2022; Akbarinia, Morgenstern, & Gegenfurtner, 2023) 和基于变压器的视觉模型 (Akbarinia et al., 2023; Cai et al., 2025) 的 CSF，但尚未有人研究聊天-视觉-语言模型（CVLMs）——能够通过生成响应集成视觉和语言的多模态系统。

传统的深度网络中 CSF 测量方法依赖于内部表示的显式读数。一些方法使用基于分类的读数 (Akbarinia et al., 2023)，训练一个分类器来检测不同对比度的光栅，将模型的感知与分类器的性能混淆。其他方法依赖于特征空间中的欧几里得距离 (Q. Li et al., 2022) 或余弦相似性 (Cai et al., 2025)，假设对比度感知对应于嵌入差异，这施加了一种可能不自然的度量。

在此，我们提出了一种行为方法，该方法反映了人类心理物理实验。我们不依赖于内部表征，而是直接询问模型关于视觉刺激的问题。具体来说，我们向 CVLMs 展示了不同空间频率下不同对比度的结构化图像，询问它们是看起来“平坦”还是包含“图案”。通过系统地变化提示并分析响应的一致性，我们以一种自然和可解释的方式估计模型的 CSF。有关方法的示意图，请参见图 1 。我们在小型开源模型（≤ 7 B 参数）上验证了我们的方法，并计划将其应用于更大型的现有模型。

方法

刺激：为了估计 CVLMs 的 CSF，我们通过在 ATD 色彩空间的无色通道中对白噪声应用带通滤波器来生成

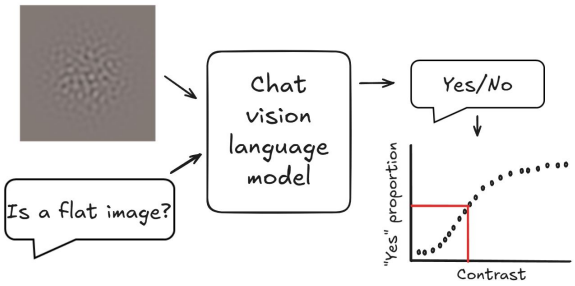

Figure 1: 方法概述：图像和问题作为输入用于聊天视觉语言模型。“是/否”答案的比例被计算为每个频率的图像对比度的函数。

256×256 图像，使用傅里叶域。这产生了感知上等效但统计上不同的刺激，较正弦光栅提高了稳健性。图像被归一化为固定的平均亮度，并被转换为 RGB 以适应模型兼容性。每幅图像跨越 4×4 视角度，与人类中央凹视力的范围一致。

任务设计：遵循人类心理物理学，每个模型都被展示一张图像和一个提示（如“< 图像 > 图像上有图案吗 ?”），要求分类为包含“图案”。为了检验提示的鲁棒性，我们系统地改变了提示，使用了“图案”的 10 个同义词，“可见”的 10 个同义词，以及 5 个改变词序的提示。每种对比-频率-提示条件被测试 10 次以生成心理物理曲线。

CSF 估计：对于每个空间频率，我们将对比度作为自变量，拟合一个逻辑函数到肯定反应的比例（即所谓的心理物理函数），这意味着模型在图像中看到一个模式。然后，我们估计模型可以 50 % 正确检测到该模式的阈值对比度。CSF 被计算为 1/阈值对比度，从而得出一个可以直接与人类 CSF 相比的敏感度曲线。

测试的模型：我们评估了多个参数数量小于或等于 70 亿的开源 CVLMs 模型：Llava1.5-7B (Liu et al., 2023)、Blip2-7B (J. Li et al., 2023)、InstructBlip-Vicuna-7B (Liu et al., 2023) 和 Qwen2.5Vl-3B (Bai et al., 2025)。每个模型获得相同的图像和提示。未来的工作将把这一方法扩展到更大的模型（Chat-GPT，Gemini-Pro）和对比训练模型（CLIP，SigLIP），以探索不同架构、训练目标和模型规模在视觉敏感性上的差异。

结果

图 2 比较了分析的聊天-视觉-语言模型的对比敏感度函数（CSFs），这是基于 25 个不同提示的平均值，与人类 CSF 的对比。人类 CSF 展示了一个典型的带通形状，在每度 4-6 个周期（cpd）附近达到峰值，然后在更高频率时下降。LLaVA-1.5-7B 是最接近人类 CSF 整体敏感度的模型，而 Qwen2.5VL-3B 尽管敏感度较低，却更好地捕捉了人类 CSF 的形状。Blip2-7B 和 InstructBlip-Vicuna-7B 得到的 CSF 平坦得多，缺乏明显的峰值。总

www.xueshuxiangzi.com

体而言，大多数模型在人类无法达到的高空间频率下表现出更高的敏感度，这表明它们专注于细节。

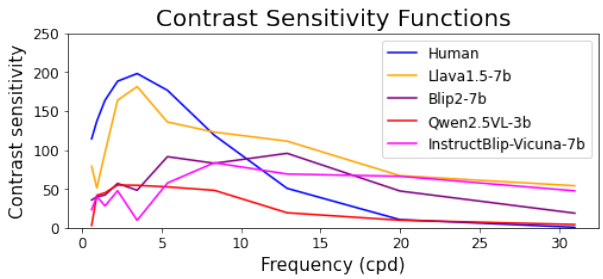

Figure 2: CSFs 结果：人体和 Chat-VL 模型在 25 个提示上的平均对比灵敏度。

表中 1 的指标证实了定性观察。Qwen2.5VL-3B 与人类 CSF 具有最高的形状 Pearson 相关性，而 LLaVA-1.5-7B 具有最低的 RMSE，表明其绝对值接近。Blip2-7B 和 InstructBlip-Vicuna-7B 在形状和幅度对齐方面表现不佳。

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: 跨模型的 CSF 相似性度量：在人类 CSF 和不同的 Chat-VL 模型 CSF 之间的 Pearson 相关性和 RMSE，以上述 25 个提示的平均值。

之前的结果通过平均 25 个提示的模型 CSF 得出，但这忽略了提示的变动性，而提示变动性可以显著影响响应。更准确的分析方法是计算每个特定提示的 CSF 的相关性和 RMSE，然后报告平均值和标准差，以评估人类 CSF 的一致性和响应的稳定性。

表 2 展示了这些结果。关于相关性，Blip2-7B 和 InstructBlip-Vicuna-7B 表现出弱正相关，表明与人类趋势的对齐程度有限。LLaVA-1.5-7B 和 Qwen2.5VL-3B 表现出负相关，表明与人类 CSF 存在偏差，尽管在提示上更一致，这反映在较低的标准差中。相比之下，Blip2-7B 表现出较高的变异性，表明对提示有很强的依赖性。RMSE 结果也呈现出类似的模式。InstructBlip-Vicuna-7B 和 Qwen2.5VL-3B 具有最低的 RMSE，表明其值更接近人类感知。LLaVA-1.5-7B 的 RMSE 最高，而 Blip2-7B 再次表现出最大的变异性。总体而言，虽然有些模型在某些情况下可能类似于人类的 CSFs，但其对提示措辞的强烈依赖引发了对一致性和稳健性的担忧。

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: 跨模型和提示集的 CSF 相似性指标：每个模型的 CSF 与人类 CSF 在 25 个不同提示下的皮尔森相关系数和 RMSE 的均值和标准差（std）。

结论

在这项工作中，我们引入了一种新的方法，使用受心理物理学启发的框架来评估基于聊天的视觉语言模型的对比敏感性。我们的研究发现，尽管这些模型对视觉对比度表现出不同程度的敏感性，但没有一个能够完全再现人类对比敏感性功能（CSF）的形状或稳定性。

在被分析的模型中，LLaVA-1.5-7B 表现出很高的整体敏感性，但显示出显著的提示依赖性，导致估计不一致。Qwen2.5VL-3B 虽然绝对而言不那么敏感，但在不同的提示下表现出更稳定的 CSF 趋势。另一方面，Blip2-7B 和 InstructBlip-Vicuna-7B 在形状对齐和稳定性方面都存在问题，突显了这些模型在处理基本感知信息方面的局限性。提示的可变性在模型性能中起到关键作用，特别是对于像 Blip2-7B 这样的模型，其响应在措辞略有变化时波动很大。这种敏感性引发了对其内部一致性和在需要细粒度视觉理解任务中的可靠性的担忧。

总体而言，这项工作为系统地探究多模态模型中的低级视觉属性提供了基础。它强调了开发提示不变评估技术的重要性，并鼓励进一步研究将模型感知与人类视觉对齐——不仅是在任务表现上，还在基本的感知过程中。

致谢

这项工作部分由 MCIN/AEI/FEDER/UE 在 PID2020-118071GB-I00 和 PID2023-152133NB-I00 资助下支持，部分由西班牙 MIU 在资助号 FPU21/02256 下支持，部分由瓦伦西亚自治区政府 GV/2021/074 和 CIPROM/2021/056 项目支持，并由 BBVA 基金会科学计划资助支持：数学、统计、计算机科学和人工智能（VIS4NN）。部分计算机资源由 Artemisa 提供，其由欧盟的 ERDF 通过粒子物理研究所 IFIC（CSIC-UV）资助。

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.

www.xueshuxiangzi.com