Dissociating spatial frequency reliance from adversarial robustness advantages in neurally guided deep convolutional neural networks

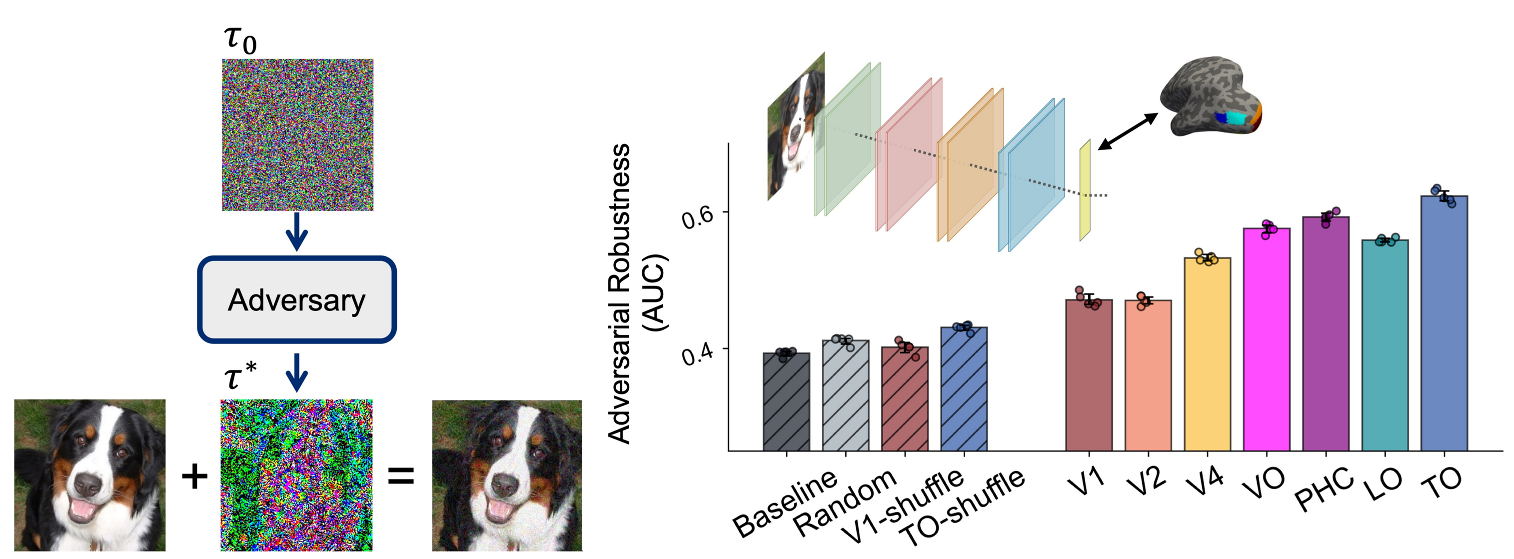
Zhenan Shao[1, 2*], Tianyu Ren[3,4], Chengxiao Wang[4], Leyla Isik[1], Diane M. Beck[2]
[1]Dept of Cognitive Science, JHU; [2]Dept of Psychology, Beckman Institute, UIUC;
[3]Dept of Civil and Environmental Engineering, UIUC; [4]Dept of Computer Science, UIUC; *Contact: zshao19@jh.edu

JOHNS HOPKINS UNIVERSITY

INTRODUCTION

Deep convolutional neural networks (DCNNs) can rival humans on many visual recognition tasks, yet they remain strikingly vulnerable to near-imperceptible perturbations generated by adversarial attacks[1].

Neural guidance (NG, aligning DCNN representations with human visual cortex) improves adversarial robustness, especially when guided by higher-order ventral-stream regions[2]. However, it remains unclear what property of the neural representation drives this benefit.

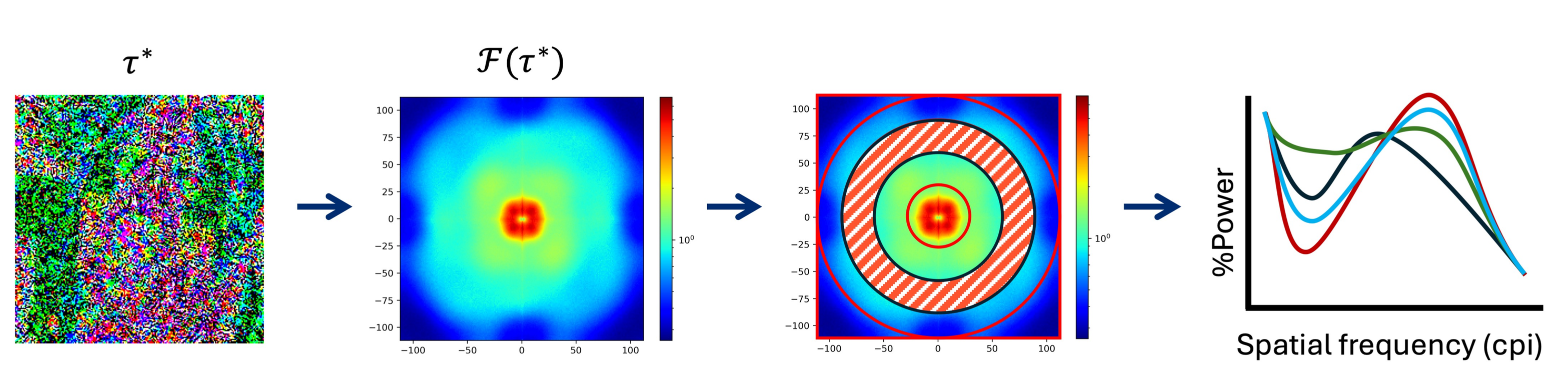


A leading hypothesis is that neural guidance steers models towards low spatial frequencies (LSF), which carry coarse, global image structure[3]. However, human object recognition has recently been linked to a narrow mid-frequency human channel[4], and prior LSF-based manipulations may have partially preserved this band.

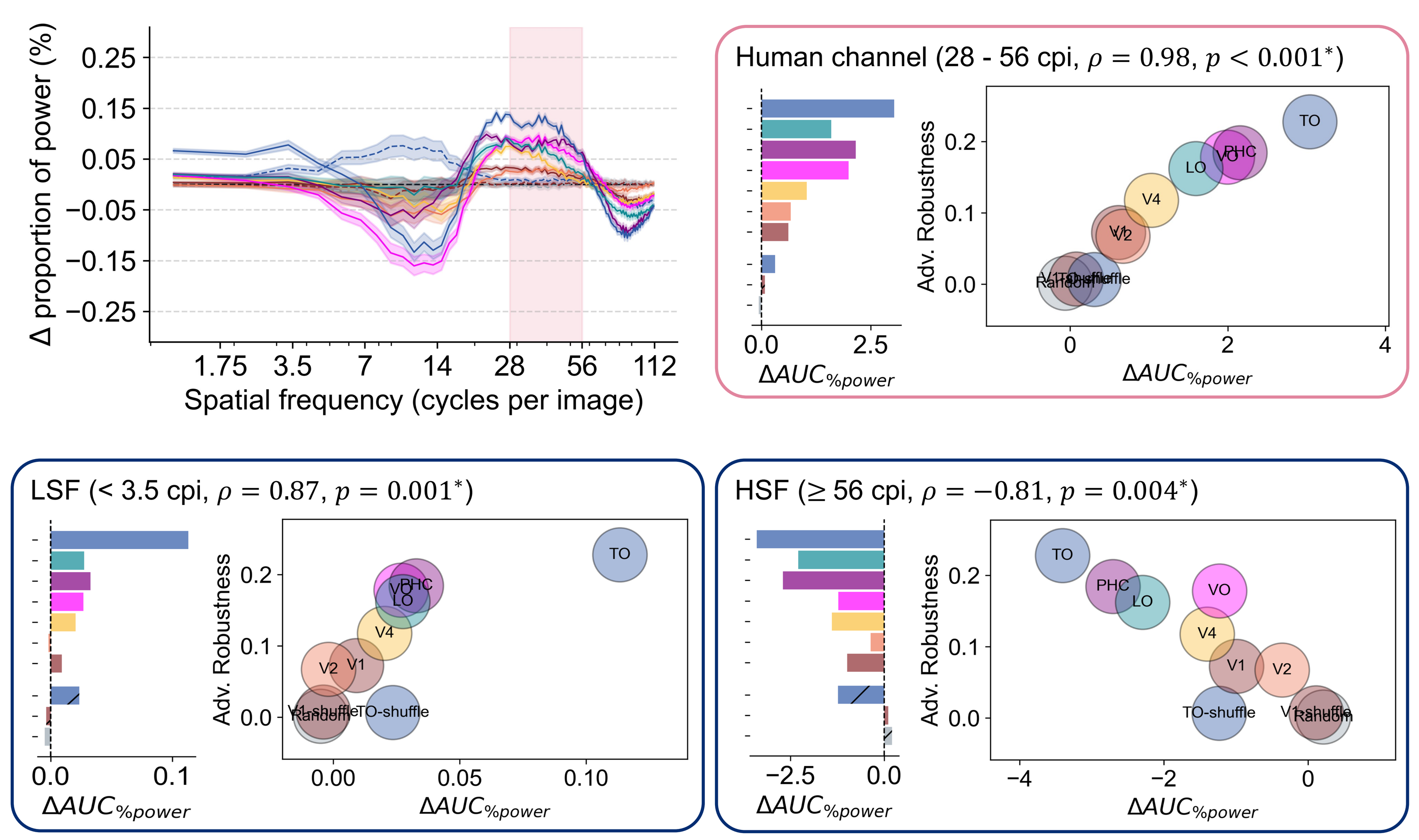
Question: Does neural guidance improve adversarial robustness by steering models towards the LSF or the human channel?

NEURAL GUIDANCE INCREASES RELIANCE ON BOTH LSF AND THE HUMAN CHANNEL

Probing SF reliance using adversarial perturbations[3]: we computed perturbations' Fourier power spectrum, summed energy within concentric radial bins, $E(r) = \sum_{|(u,v)|=r} P(u,v)$, and then normalized the result, $\tilde{E}(r) = \frac{E(r)}{\sum_r E(r)}$, to estimate the relative reliance on each frequency band (cycles/image, cpi). We then compared these spectra across the NG models.

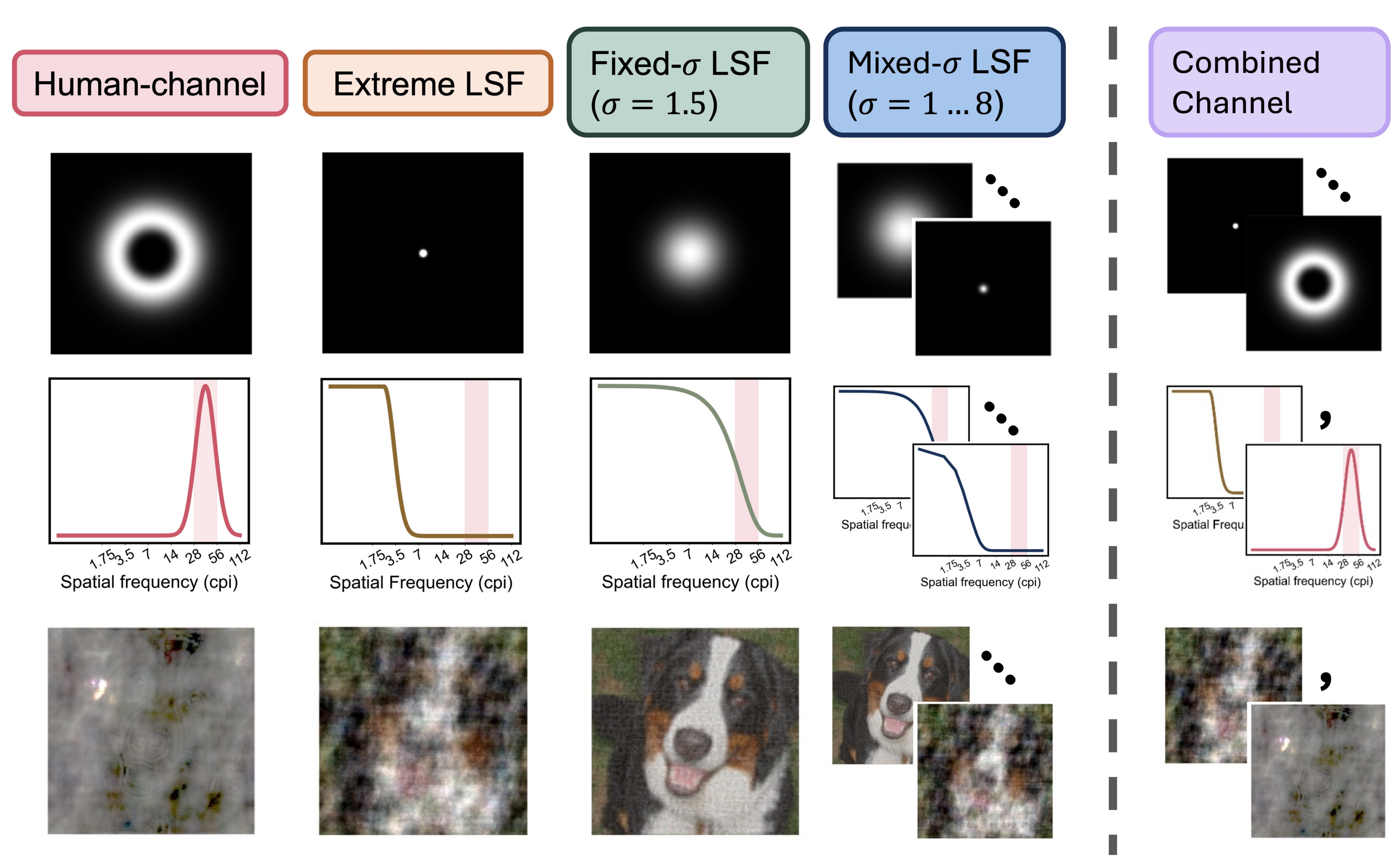


Results: Neural guidance steered models' reliance towards both LSF (< 3.5 cpi) and the human channel[4] (28–56 cpi), and these shifts tracked robustness gains across NG-models, suggesting both bands could be candidate spectral components inherited from neural representations that drives NG models' robustness gains.



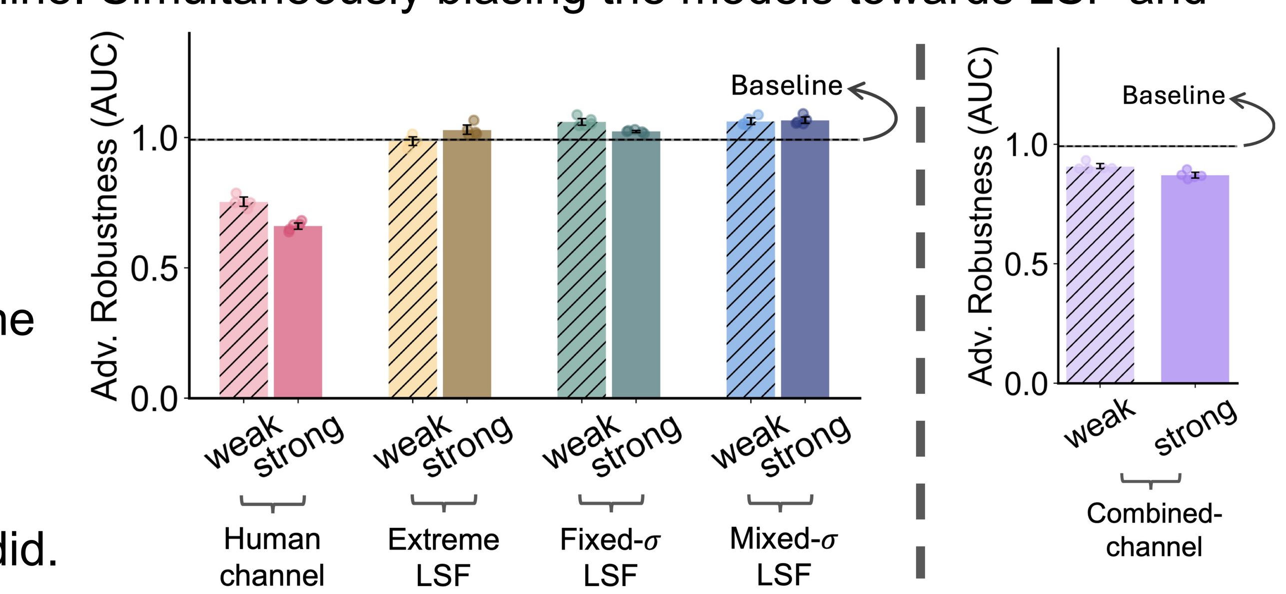
DIRECT SF BIASING DISSOCIATES HUMAN-CHANNEL AND LSF EFFECTS

Spatial frequency biasing: We trained models with *band-selective phase-scrambled* images to force reliance on specific SF bands: the human channel[4], extreme LSF (< 3.5 cpi), or fixed[3]-σ / mixed[5]-σ LSF ranges. Proportion of SF-manipulated images were varied. Robustness was measured using PGD[6] robust AUC.



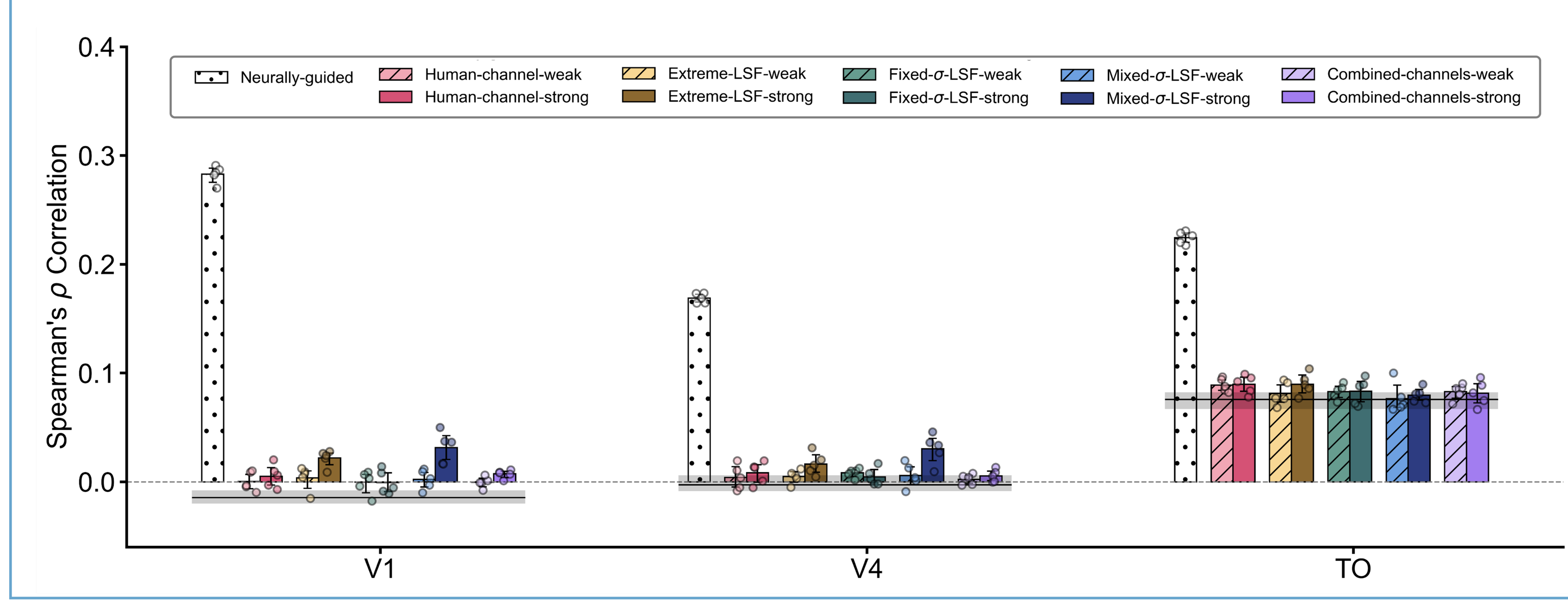
Results: Human-channel bias was harmful. LSF-biased models performed near or only slightly above baseline. Simultaneously biasing the models towards LSF and the human channel still decreased robustness.

Therefore, spectrally biasing models towards the human channel or the LSF, or both, did not reproduce robustness benefits that aligning to neural representations did.



SF BIAS DOES NOT RECOVER HUMAN-LIKE REPRESENTATIONAL GEOMETRY

Using RSA[7], we compared SF-biased models and NG models with human fMRI recordings[8]. SF-biased models, including the combined-channel model, showed poor representational similarity with the ROIs, remaining far from the NG models.



TAKEAWAYS

1. Neural guidance increased reliance on both LSF and the human channel, but explicitly biasing models towards these bands, alone or in combination, failed to reproduce the robustness gains or human-like representational geometry.
2. Spatial-frequency reliance shifts are emerging properties of learning more brain-like representations, but not a sufficient explanation for robustness.

ACKNOWLEDGEMENTS

This work was supported by NIH R01MH132826 awarded to L.I.

REFERENCES

[1][Szegedy et al., arXiv., 2014] [2][Shao et al., arXiv, 2025] [3][Li et al., Plos. Comp. Bio., 2023] [4][Subramanian et al., Neurips, 2023] [5][Jang & Tong, Nat. Comm., 2024] [6][Madry et al., ICLR, 2018] [7][Kriegeskorte, Frontiers in Sys. Neuro., 2008] [8][Allen et al., Nat. Neurosci., 2022]

Paper