
Dense Jacobian Transport¹

*Changes Model-Brain RSA in a Prompt-
and Depth-Dependent Manner*

Frank Peterlein
Independent Researcher

With
Apart Research

Abstract

Representational similarity analysis (RSA) can reveal shared geometry without identifying a shared mechanism. We tested whether dense Jacobian transport of Qwen3-4B states improves alignment with human visual-cortical activity. Across 515 Natural Scenes Dataset images, 2,578 COCO captions, eight participants, three higher-visual regions, and raw-caption layers 18 and 27, transport reduced mean RSA from .09782 to .08558 for every participant (mean difference=-.01224; Cohen's d =-3.07; exact two-sided sign-flip p =.0078125). A post-result extension frozen before new extraction added raw-caption layers 9 and 34 and imagery-prompt layers 9, 18, 27, and 34. Fifteen of 18 new region-level contrasts were negative, all three positive estimates occurred under imagery at layer 18, and 17 of 18 rejected after Benjamini-Hochberg correction. Full-rank but strongly anisotropic Jacobians and a no-final-normalization control indicate geometric reweighting rather than hard rank loss or one uniform normalization artifact. This is a controlled boundary test of dense transport, not a test of the formal sparse, non-negative J-space projection or a biological global workspace.

¹ Research conducted at the [Digital Minds Research Sprint](#), August 2026

1. Introduction

Language-model representations correlate with neural responses to natural stimuli, but a shared representational geometry does not imply a shared computation [1–3]. The new Jacobian Lens (J-Lens) identifies directions that can be transported toward a model’s output and motivates a workspace-like interpretation of a sparse, verbalizable subspace [4]. This creates a timely empirical question: do J-Lens-derived representations align differently with human brain activity than ordinary residual-stream states?

We answer a narrower question than the biological-workspace claim. Our tested representation is the normalized dense transport $J_l h_l$, not the sparse non-negative projection $\Pi_J(h_l)$ defined in the source work. Dense transport is therefore a control and diagnostic: it asks whether emphasizing output-relevant directions generically improves model–brain RSA. It cannot establish that a brain region implements J-space or a global neuronal workspace.

Contributions. We provide a reproducible boundary test with prospectively separated confirmatory and exploratory stages:

1. A preregistered subject-level RSA comparison using shared NSD images, caption-derived model features, and exact randomization inference.
2. A frozen prompt-by-depth extension across three higher-visual regions, with a declared 18-cell multiplicity family.
3. Spectral and normalization diagnostics that distinguish rank loss from anisotropic reweighting and clarify what the dense transform can support.

2. Related Work

RSA compares the rank geometry of dissimilarity matrices and is widely used for model–brain comparisons [1]. The Natural Scenes Dataset (NSD) provides repeated 7T fMRI responses to thousands of natural images [2]. Doerig et al. showed that language-model embeddings can predict or align with human visual representations, providing the immediate pipeline we adapt [3]. J-Lens instead studies model-internal geometry through layer-to-output Jacobians and a sparse non-negative, vocabulary-readable factorization [4]. Global-workspace theories concern access, reportability, and widespread availability [5]; our analysis only measures representational correspondence. Benjamini-Hochberg false-discovery-rate control is used for the declared exploratory family [6].

3. Methods

3.1 Data and neural representations

We analyzed 515 NSD images shared across all eight participants and 2,578 associated COCO captions [7]. For each participant we used preprocessed response patterns from bilateral ventral, lateral, and parietal higher-visual regions. Within each participant and region, image-by-voxel responses were converted to correlation-distance representational dissimilarity matrices (RDMs). Every participant contributed all 515 images and three finite, nonconstant RDMs; no participant was excluded after frozen quality control.

3.2 Model representations and prompts

Qwen3-4B was loaded at revision 1cfa9a7208912126459214e8b04321603b3df60c with matching public J-Lens assets. Each caption was encoded separately; the last attended non-terminal-special-token state was read out and caption features were averaged within image. The baseline was the model’s learned final-RMSNorm applied to h_l ; the dense condition applied the same readout to $J_l h_l$. The preregistered analysis used raw captions at layers 18 and 27. The frozen extension added raw-caption layers 9 and 34 and imagery prompting at layers 9, 18, 27, and 34 with the exact template "Imagine the visual scene described by this caption: <caption>". Correlation-distance model RDMs were computed from image features.

3.3 RSA and inference

For each participant, region, and representation, we computed Pearson correlation between the upper triangles of model and neural RDMs. The primary participant statistic averaged six layer-by-region contrasts

(layers 18/27 x ventral/lateral/parietal). The paired contrast was $\Delta = r_{\text{transport}} - r_{\text{raw}}$; negative values favor ordinary residual geometry. The confirmatory test was an exact two-sided sign-flip test over eight participants. The extension family comprised 18 newly extracted cells: raw layers 9/34 plus imagery layers 9/18/27/34, each crossed with three regions. Benjamini-Hochberg correction was applied once across that family. Extension choices were frozen after the primary result but before new extraction and remain exploratory. The 132,355 image-pair dissimilarities were not treated as independent observations.

4. Results

4.1 Confirmatory result

The residual baseline aligned positively with higher-visual cortex (mean RSA $r = .09782$), but dense transport reduced the mean to $r = .08558$. Every participant showed a negative aggregate difference (mean $\Delta = -.01224$; SEM = .00141; Cohen's $d_z = -3.07$; exact two-sided sign-flip $p = .0078125$; Figure 1A). Five of six layer-by-region estimates were negative and survived FDR correction; the exception was a small, nonsignificant lateral-cortex estimate at layer 27 ($\Delta = +.00302$). Thus the predicted generic improvement was not observed for the preregistered raw-caption, two-layer, three-region aggregate.

4.2 Prompt-by-depth extension

The depth profile was configuration-dependent. Averaged across the three regions, raw-caption contrasts were $-.07157$, $-.02134$, $-.00315$, and $-.02510$ at layers 9, 18, 27, and 34. Imagery-prompt contrasts were $-.02968$, $+.00430$, $-.01066$, and $-.02832$ (Figure 1B). Within the 18-cell new-condition family, 15 estimates were negative; all three positive estimates occurred under imagery at layer 18. Seventeen rejected after correction. The sole non-rejection was parietal imagery-layer-18 ($\Delta = +.00286$; exact and adjusted $p = .05469$). Because this family was frozen only after the primary result, its corrected p-values quantify within-extension evidence but do not make the pattern confirmatory.

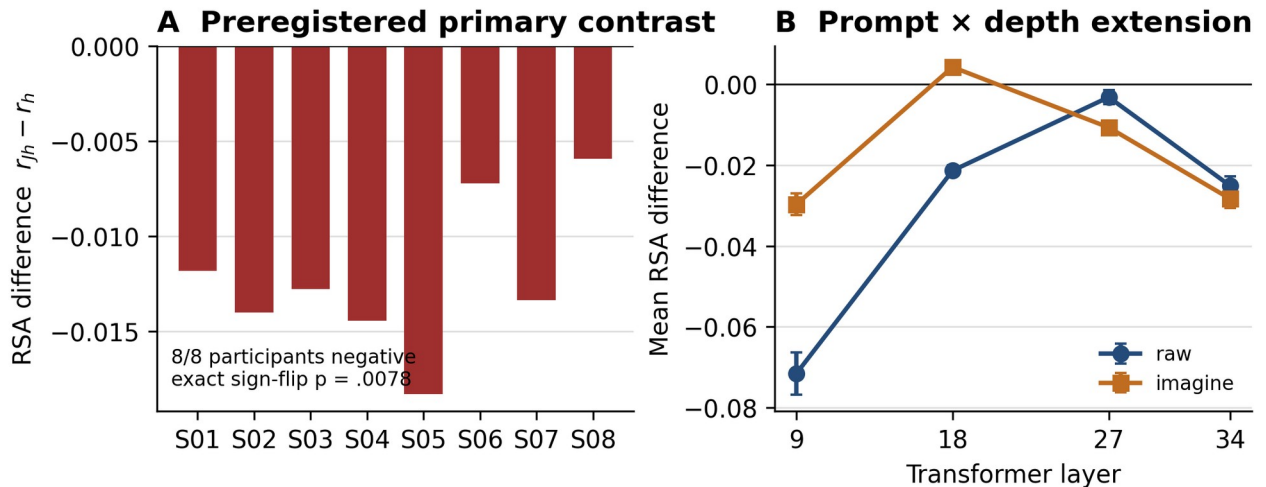


Figure 1. Dense-transport difference in RSA. (A) The preregistered raw-caption aggregate across layers 18/27 and three higher-visual regions was negative in every participant. (B) The descriptive prompt-by-depth trajectory varied by configuration; error bars are participant-level SEM after averaging the three regions.

4.3 Diagnostics

The analyzed $2,560 \times 2,560$ Jacobians were numerically full rank, ruling out simple hard rank collapse. Effective rank increased from 590 (layer 9) and 1,157 (layer 18) to 2,226 (layer 27) and 2,483 (layer 34), while condition numbers fell from approximately 1.80×10^6 and 2.43×10^6 to 6.25×10^3 and 2.32×10^3 . Raw-versus-transported model-RDM correlations likewise rose with depth: .697, .719, .848, and .961 for raw captions; .789, .803, .944, and .978 for imagery prompts. Without the final normalization, layer-18 contrasts remained negative in all three regions ($-.01557$ lateral, $-.00695$ parietal, $-.01858$ ventral), whereas all three layer-27 contrasts were positive ($+.00845$, $+.00201$, $+.00910$). The effect therefore reflects strong anisotropic reweighting and a layer-dependent normalization interaction, not one uniform normalization artifact.

5. Discussion and Limitations

Dense J-Lens transport does not provide a general shortcut to stronger brain alignment. Its effect depends on layer and prompt: early and middle Jacobians substantially reweight geometry, late transport increasingly preserves the baseline RDM, and imagery reverses the aggregate sign only at layer 18. That isolated positive pattern is a targeted replication hypothesis, not a discovery of a workspace region.

Limitations. The participant sample is fixed at eight. Image-evoked fMRI is compared with human captions, which are already semantic compressions of scenes; lower-level visual interpretation is therefore limited. We tested one post-trained model/lens, one pooling rule, three higher-visual regions, and no early-visual control. The extension followed the primary result. RSA establishes shared dissimilarity structure, not shared features, computation, causality, consciousness, or biological implementation. Most importantly, dense J_lh_l is not the source paper's sparse, non-negative, vocabulary-interpretable J-space projection.

Future work. A prospectively frozen diagnostic compares actual transport with the singular-value-flattened polar factor and 63 spectrum-preserving coordinate permutations; it is not reported as a result until its numerical gates pass. The decisive J-space study must compare the formal sparse projection with its complement, a variance-matched random subspace, and the full residual state, while adding early visual cortex, replicating imagery-layer-18, and reporting encoding models alongside RSA.

6. Conclusion

Dense Jacobian transport reliably changed model-brain RSA, but usually reduced it and did so in a prompt- and depth-dependent pattern. The result rules out a simple embedding-swap improvement for this pinned configuration and narrows the next mechanistic tests. Formal sparse J-space, matched subspace controls, and regional specificity are required before any neural-biological conclusion.

Code and Data

Code and reproducibility materials: <https://github.com/frnkptrln/jlens-model-brain-rsa>. The repository contains pinned configuration, preprocessing and RSA code, tests, manifests, participant-level tables, frozen analysis documents, and figure scripts. Natural Scenes Dataset: <https://naturalscenesdataset.org/>. Raw NSD data remain governed by NSD access terms and are not redistributed.

References

- [1] Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4. <https://doi.org/10.3389/neuro.06.004.2008>
- [2] Allen, E. J., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 116–126. <https://doi.org/10.1038/s41593-021-00962-x>
- [3] Doerig, A., et al. (2025). High-level visual representations in the human brain are aligned with large language models. *Nature Machine Intelligence*, 7, 1220–1234. <https://doi.org/10.1038/s42256-025-01072-0>
- [4] Gurnee, W., Sofroniew, N., Pearce, A., et al. (2026). Verbalizable Representations Form a Global Workspace in Language Models. *Transformer Circuits*. <https://transformer-circuits.pub/2026/workspace/index.html>
- [5] Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. *Cognition*, 79, 1–37. [https://doi.org/10.1016/S0010-0277\(00\)00123-2](https://doi.org/10.1016/S0010-0277(00)00123-2)
- [6] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [7] Lin, T.-Y., et al. (2014). Microsoft COCO: Common Objects in Context. *European Conference on Computer Vision*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Appendix A. Confirmatory, Exploratory, and Diagnostic Conditions

Stage	Prompt	Layer(s)	ROIs	Inference
Preregistered	raw caption	18/27	ventral/lateral/ parietal	one exact two-sided sign-flip test
Post-result frozen extension	raw 9/34 + imagery 9/18/27/34	9/18/27/34	3 region aggregates	BH across 18 new cells
Frozen diagnostics	raw + imagery	9/18/27/34	model RDMs + same regions	spectra, RDM, no- final-norm

LLM Usage Statement

OpenAI Codex was used to implement and test analysis code, inspect run artifacts, generate figures, and edit this manuscript. Claude was used for strategic critique of the study framing. All numerical claims were checked against saved tables and manifests; the human author remains responsible for scientific decisions, interpretation, authorship, and final verification. No language-model output was used as neural data, and no AI system is listed as an author.