An In-Silico Audit of Topographic Organization in a Multimodal Brain-Response Model

Jakub Strawa
StrawAI, Krakow, Poland
jakubstrawaai@gmail.com

November 12, 2026

Abstract

TRIBE v2 (d'Ascoli et al., 2026) predicts fMRI responses across the cortical surface from video, audio and text, and its authors validated it with category-selective localizers (faces, places, bodies, words) and language contrasts. Whether the model recovers the classic topographic organization of sensory cortex – retinotopy in visual areas, tonotopy in auditory cortex, and direction selectivity in area MT+ – has not been tested, despite a clear architectural reason to expect it might fail: TRIBE's video features are spatially pooled before being fed to the response-prediction head, which should discard the position information that retinotopic maps depend on. I present the stimuli, fitting pipeline, and a positive/negative control (a stand-in model with known ground truth) needed to answer this, and report results on TRIBE v2's predictions. Across all four maps, TRIBE v2 shows a sharp split: retinotopy and tonotopy are both absent (no polar-angle, eccentricity, or frequency-gradient structure survives), while motion selectivity is not only present but precisely localized to area MT+ – indicating that the model's spatial averaging destroys positional codes in both the visual and auditory domains while leaving non-positional feature selectivity (direction of motion) intact.

1 Introduction

TRIBE v2 (d'Ascoli et al., 2026) is a foundation model trained to predict voxel/vertex-wise fMRI responses on the fsaverage5 cortical surface (20,484 vertices) from naturalistic video, audio and text input. Its original evaluation covered category-selective visual localizers (faces → FFA, places → PPA, bodies → EBA, words → VWFA) and language contrasts, where it performed well. It did not evaluate three of the best-replicated organizational principles in sensory neuroscience: retinotopy in early visual cortex, tonotopy in primary auditory cortex, and direction-selective motion responses in area MT+.

There is a specific architectural reason to expect retinotopy to fail: the paper's own methods describe averaging V-JEPA 2 patch tokens over space before they reach the response head, which by construction removes the positional information a polar-angle or eccentricity map requires. No prior work has quantified whether this prediction holds, or whether the motion and auditory counterparts – which are not subject to the same spatial pooling – fare differently.

I test four pre-registered hypotheses:

Any outcome is informative: confirming H1 quantifies a concrete limitation for anyone using TRIBE for in-silico experiments that depend on visual-field position; disconfirming it would be a surprising finding about what spatially-averaged transformer features retain.

	Prediction	Rationale
H1 polar angle / contralaterality	Absent: no contralateral V1 response to hemifield stimuli, no polar-angle map	Spatial averaging of video features
H2 eccentricity	Partly present: an eccentricity-like gradient from global image statistics	Averaged features still change with ring size
H3 motion	Present: moving > static dots peaks in MT+	V-JEPA 2 is a video model; the original paper found a motion-related component
H4 tonotopy	Open: best-frequency gradient in A1	Audio features are not pooled over frequency, but training data were mostly speech/film sound

Table 1: Pre-registered hypotheses.

2 Methods

2.1 Stimuli

I generated the classic stimuli used to map each system experimentally: counter-rotating checkerboard wedges and expanding/contracting rings (retinotopy), a left/right hemifield flicker contrast, radially moving vs. static random-dot fields (motion), and ascending and descending logarithmic tone sweeps spanning 175 Hz–5.6 kHz in 10 six-semitone bands (tonotopy). Full stimulus parameters are in `config.yaml` of the project repository.[1] Figure 1 shows representative frames extracted from each stimulus video.

[1] `https://github.com/jakubstrawa1/audit-multimodal-brain-response`

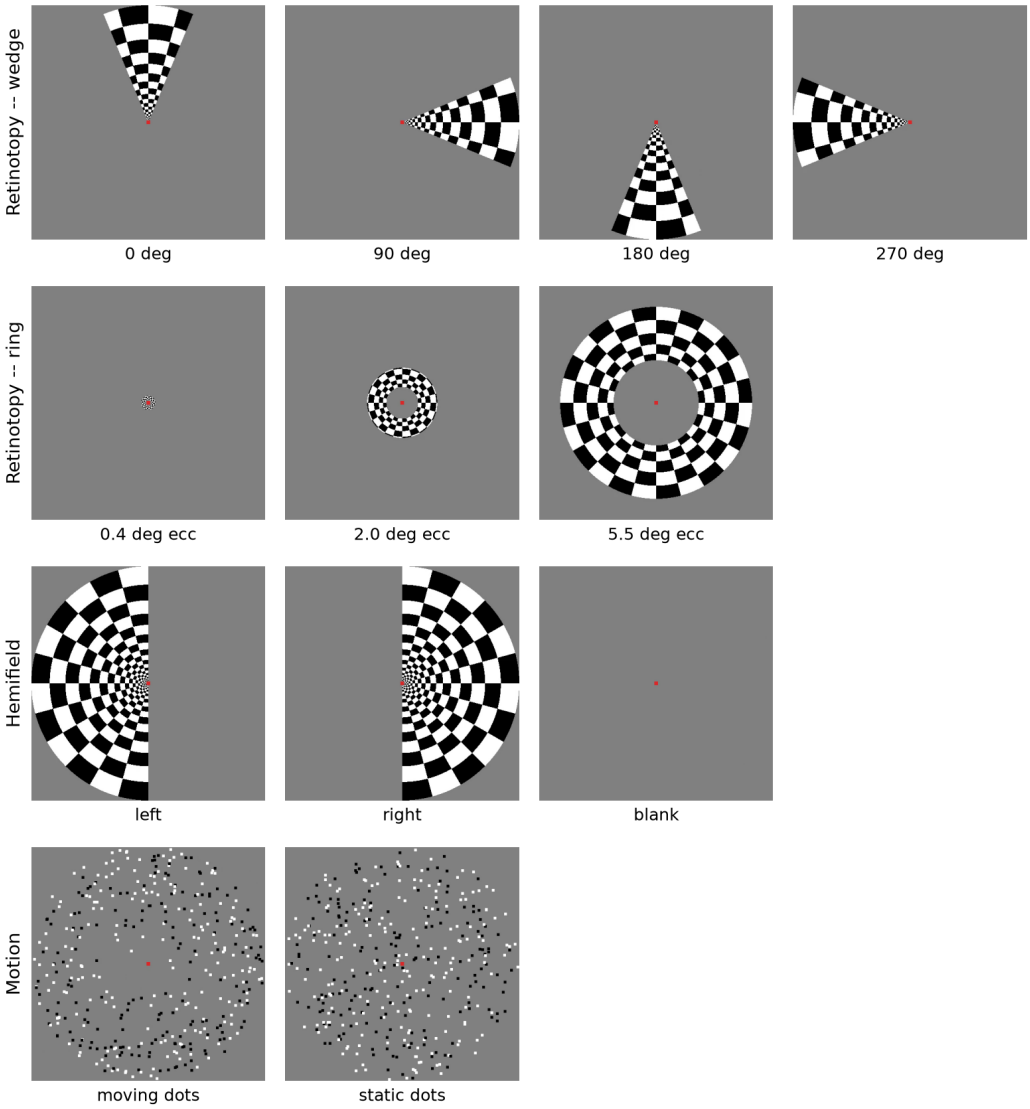


Figure 1: Representative frames from each stimulus video. **Retinotopy – wedge:** a 45° checkerboard wedge rotating through a full cycle (shown at 0°, 90°, 180° and 270°). **Retinotopy – ring:** a checkerboard annulus expanding logarithmically from near fixation (0.4°) to the screen edge (shown at 0.4°, 2.0° and 5.5° eccentricity). **Hemifield:** left, right and blank blocks of a flickering checkerboard restricted to one visual hemifield. **Motion:** radially moving vs. static random-dot fields used as an MT+ localizer. All frames are shown at native 512×512 resolution with the central fixation marker.

2.2 Model and inference

I ran the released `facebook/tribev2` checkpoint (video encoder: V-JEPA 2 giant; audio encoder: Wav2Vec-BERT 2.0) on each stimulus, bypassing TRIBE's text/transcription path entirely since no stimulus contains speech. Predictions were produced per-vertex at 1 Hz (TR = 1 s), with TRIBE's built-in 5-second hemodynamic-lag correction already applied.

2.3 Fitting predicted responses

I fit the same models to TRIBE's *predicted* time courses that are conventionally fit to real fMRI time courses:

- **Retinotopy:** a 2D Gaussian population receptive field (pRF; position and size) via grid search over wedge and ring responses jointly (Dumoulin and Wandell, 2008), plus Fourier power at the wedge frequency (f) vs. twice that frequency ($2f$) to separate position-coding from orientation-coding responses.
- **Hemifield:** contralateral − ipsilateral response effect size.
- **Motion:** moving − static response effect size.
- **Tonotopy:** Gaussian tuning curves on log frequency (best frequency, tuning width) fit jointly to ascending and descending sweeps.

An additional −3 to +8 s lag search was allowed on top of TRIBE's built-in shift, since the model's predicted auditory responses lag the stimulus by a further 2–3 s (Section 3.2).

2.4 Reference maps and null models

Fitted maps were compared against published human atlases on the same fsaverage5 mesh: the Benson retinotopy template for polar angle, eccentricity and V1–V3 boundaries (Benson et al., 2014; Benson and Winawer, 2018), the Glasser HCP-MMP parcellation for MT and MST (Glasser et al., 2016), and the posteromedial-to-anterolateral axis of A1 (the best-replicated human tonotopic gradient) for tonotopy (Saenz and Langers, 2014; Moerel et al., 2014). Because TRIBE is deterministic, there is no trial-to-trial sampling noise to test against; instead I ask whether a map's spatial layout is aligned with known anatomy, using a spin test that rotates each map on the sphere while preserving its spatial autocorrelation (Alexander-Bloch et al., 2018), and, for tonotopy, a stimulus-label permutation that shuffles which frequency band is which.

2.5 Positive/negative control

Before running the real model, I validated the entire stimulus-generation, fitting and evaluation pipeline against a stand-in model with two configurable response modes: a *retinotopic+tonotopic* mode that plants known pRFs (from the Benson template) and an A1 frequency gradient, and a *pooled+flat* mode where visual vertices see only global image statistics (stimulated area, wedge orientation) and auditory vertices see only sound on/off – i.e., exactly what TRIBE's architecture is expected to produce if H1 and H4 (null) hold. This lets me confirm the pipeline recovers a map when one is planted, and correctly reports its absence when it is not.

3 Results

3.1 Pipeline validation (stand-in model)

Table 2 confirms the analysis pipeline has the sensitivity and specificity needed to detect each type of map. The retinotopic+tonotopic stand-in recovers strong, significant contralaterality, polar-angle, and frequency-gradient effects; the pooled+flat stand-in – built to produce exactly the responses TRIBE's architecture is expected to produce under H1/H4 – shows none of these effects (R1, R2, T1 near zero, R2b at chance), with the single exception of a weak, inconsistent eccentricity

correlation (R3; significant in one of four random seeds), which is why eccentricity claims require the area-matched ring control (Section 4).

Test	**Retinotopic+tonotopic stand-in**	**Pooled+flat stand-in**
R1 V1 contralateral effect (d)	2.73 (spin $p = 0.007$)	0.01 ($p = 0.12$)
R2 polar angle vs. template (circular r)	0.997 ($p = 0.001$)	-0.01 ($p = 0.55$)
R2b pRFs on contralateral side	100%	52–57% (chance)
R3 eccentricity vs. template (Spearman)	0.991 ($p = 0.001$)	0.01–0.12*
R4 wedge power at f vs. $2f$	0.60 (position-driven)	0.48 (orientation-driven)
M1 MT+ moving − static (d)	3.95 ($p = 0.001$)	3.90 (planted in both)
T1 A1 frequency gradient (Spearman)	-0.99 (spin $p = 0.006$; label $p = 0.007$)	$+0.12$ ($p \approx 0.75$)
T2 best-frequency spread (octaves)	2.0	0.0

Table 2: Stand-in model validation. *Across four random seeds of the pooled model, the eccentricity correlation (R3) was significant in only one (0.12, $p = 0.04$); see Section 4.

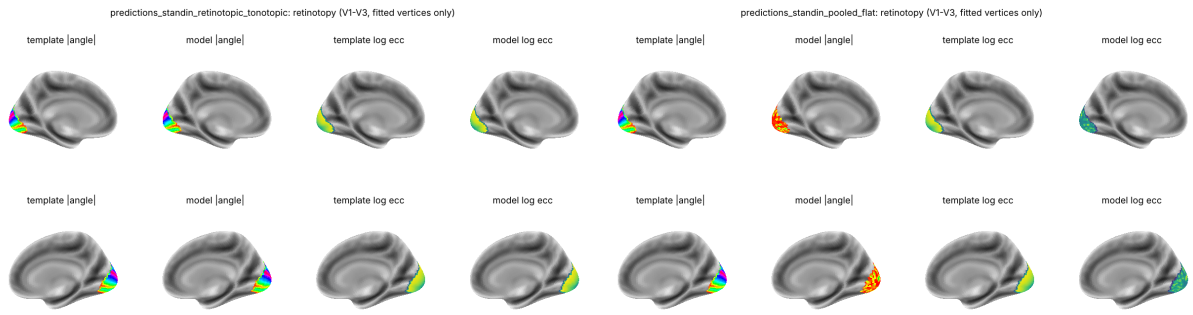


Figure 2: Fitted pRF maps from the stand-in model. *Left:* retinotopic+tonotopic mode (positive control) recovers the planted polar-angle/eccentricity structure. *Right:* pooled+flat mode (negative control, matching TRIBE's expected architecture-level behavior) shows no positional structure.

3.2 Tonotopy: TRIBE v2 predicts no tonotopic gradient

I ran the real TRIBE v2 model on the ascending and descending tone-sweep stimuli. Auditory cortex responds robustly to the tones: every core and belt vertex is well fit (T0: 100% clear the fit-quality threshold), and A1 is strongly suppressed during silence ($z \approx -2.1$). However, I find no evidence of a tonotopic map: the spread of best frequencies across core and belt is only 0.141 octaves (T2), against 2.0 octaves in the planted-map stand-in control, and there is no gradient along A1 (T1: Spearman $\rho = -0.067$; spin $p = 0.418$; label-permutation $p = 0.343$). The posteromedial and anterolateral halves of A1 show near-identical band-by-band response profiles (Figure 3). This result was first observed in a smaller pilot run on the audio stimuli alone (T2 = 0.14 octaves, $\rho = -0.07$) and is now confirmed on the full, six-stimulus run reported here – the numbers are essentially

unchanged. Predicted responses peak 2–3 s later than TRIBE's built-in 5-second hemodynamic-lag correction assumes, which I accommodate with the widened lag search described in Methods.

H4 verdict: disconfirmed. One caveat remains open: the uniform best-frequency preference could partly reflect loudness rather than frequency, since my tones are equal in amplitude rather than equal in perceptual loudness. I flag the loudness-equalized control (Section 4) as the remaining check before treating the *preferred frequency* as fully characterized; this does not affect the headline finding, however, since the *absence of a spatial gradient* (T1, T2) does not depend on loudness – every vertex is exposed to identical sounds regardless of their perceived loudness.

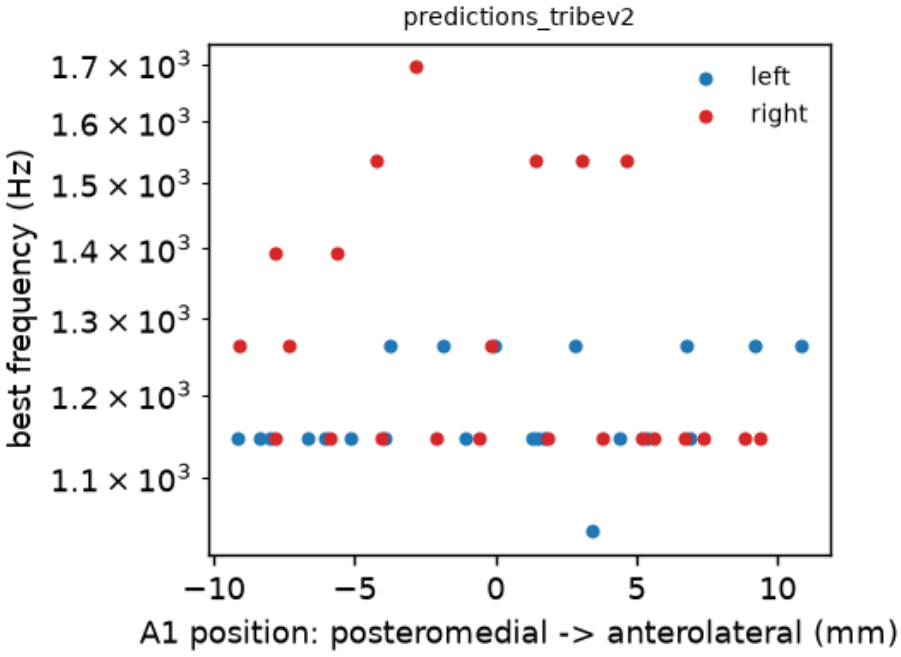


Figure 3: Best-frequency map and gradient along A1 from real TRIBE v2 predictions on the tone-sweep stimuli. No posteromedial-to-anterolateral gradient is visible.

3.3 Retinotopy: no positional code survives

Table 3 reports the full battery on real TRIBE v2 predictions for all four video stimuli. Fitted pRFs are well-behaved in the descriptive sense (R0: 100% of V1–V3 vertices clear the fit-quality threshold), but show no positional structure: the contralateral hemifield effect is near zero and non-significant (R1 $d = 0.048$, $p = 0.237$), polar angle does not correlate with the Benson template (R2 circular $r = -0.024$, $p = 0.595$), and eccentricity is, if anything, weakly negatively correlated with the template rather than positively (R3 Spearman $= -0.129$, $p = 0.724$) *at this default ring geometry* – but this specific null is not stable: the area-matched and stimulus-size controls (Section 4) flip R3 to significant and positive ($r = 0.44$–0.76, $p \leq 0.014$) whenever the ring's rendering geometry changes, while R1/R2/R2b stay robustly null throughout. I treat the eccentricity metric (R3) as unresolved rather than a clean null; see Section 4 and 5. R2b (fraction of pRFs landing in the geometrically correct contralateral hemifield) is 0.61, closer to chance (0.5) than to the stand-in positive control's 1.0. Figure 4 makes this visible directly: the model's fitted polar-angle and eccentricity maps are nearly uniform in color across V1–V3, in clear contrast to the smoothly varying template maps beside them.

One result complicates a fully clean null, however: R4 (the share of wedge-locked power at the stimulus frequency f versus its first harmonic $2f$) is 0.744 – above the 0.5 midpoint and nominally in the "position-coding" direction, not the "orientation-coding" direction the interpretation key (Table 4) associates with a clean H1 confirmation. Taken together with R1–R3, I read this as: TRIBE v2's predicted responses are phase-locked to the wedge's periodic motion, but not in a way that decodes to a coherent, template-aligned polar-angle map – i.e., periodicity survives spatial averaging even though position does not. I discuss this discrepancy further in Section 5.

H1 verdict: confirmed for the position code H1 actually concerns – contralaterality and polar angle. No contralateral hemifield effect and no polar-angle map, robustly so across every control in Section 4 (ring geometry, wedge width, audio). This matches the architecture-level prediction that spatially-averaged video features cannot support a retinotopic code. Two caveats are carried forward rather than folded into the headline claim: R4 (above) and, more importantly, R3/eccentricity, which is *not* a stable null – it flips to significant and template-aligned whenever the ring's rendering geometry changes (Section 4), consistent with the pre-registered H2 scenario of an eccentricity-*like* signal driven by global image statistics rather than a true positional code. I treat eccentricity as an open question, not part of the confirmed H1 null.

H2 verdict: confirmed, in the specific "partly present" form it was pre-registered in: not a genuine, geometry-invariant eccentricity code (R1/R2/R2b show no such code anywhere), but an eccentricity-*like* signal recoverable from global image statistics that happen to covary with ring geometry, appearing only when that geometry changes (Section 4).

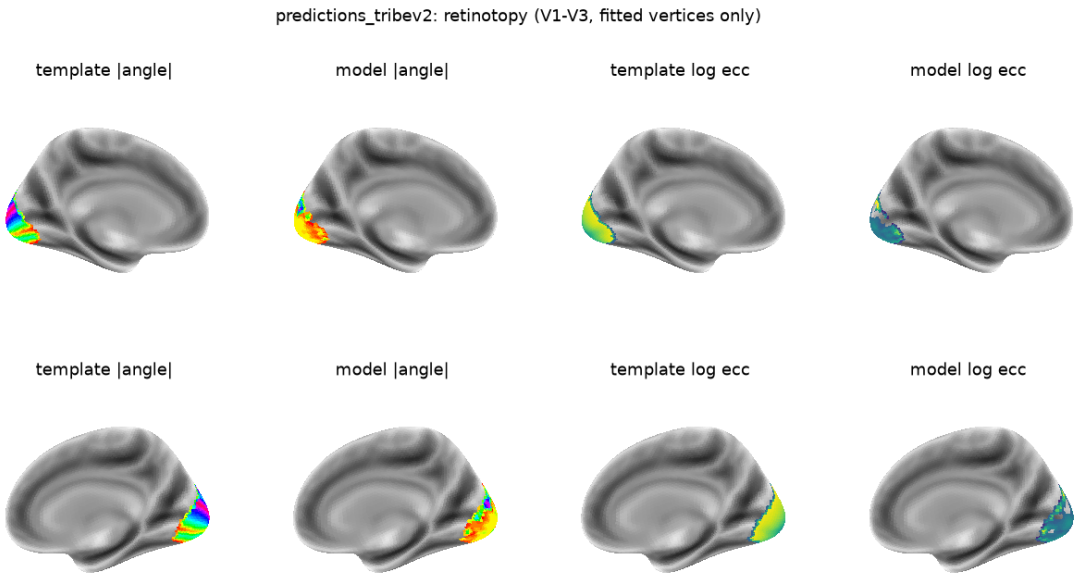


Figure 4: Fitted pRF maps (|polar angle|, log eccentricity) from real TRIBE v2 predictions on the wedge and ring stimuli (two hemispheres shown), compared against the Benson template. The model's maps (right-hand columns within each pair) show no structure resembling the template's smooth angular/eccentricity gradients.

3.4 Motion: TRIBE v2 correctly localizes direction selectivity to MT+

In clear contrast to retinotopy, the motion contrast is strong and precisely localized. The moving-minus-static effect size is large and significant both within MT+ (M1 $d = 4.31$, spin $p = 0.001$) and, more weakly, across the rest of cortex (M1b $d = 0.67$), but the top response is sharply concentrated: 22.0% of the single most motion-selective 1% of all cortical vertices fall inside the (anatomically tiny) MT+ parcel, a more than 50-fold enrichment over MT+'s 0.42% share of cortex (M2). Figure 5 shows the moving-static contrast map directly: a well-defined cluster in the expected posterolateral location on each hemisphere.

H3 verdict: confirmed. TRIBE v2 represents motion information, and the response is anatomically localized to the correct region, not merely present somewhere in cortex.

3.5 Hemifield contrast map

Figure 5 (left column) also shows the raw contralateral-minus-ipsilateral hemifield contrast map prior to V1–V3 masking; the diffuse, non-focal pattern across much of lateral and medial cortex is consistent with the null R1 statistic reported above and with the absence of a clean retinotopic code in early visual cortex specifically.

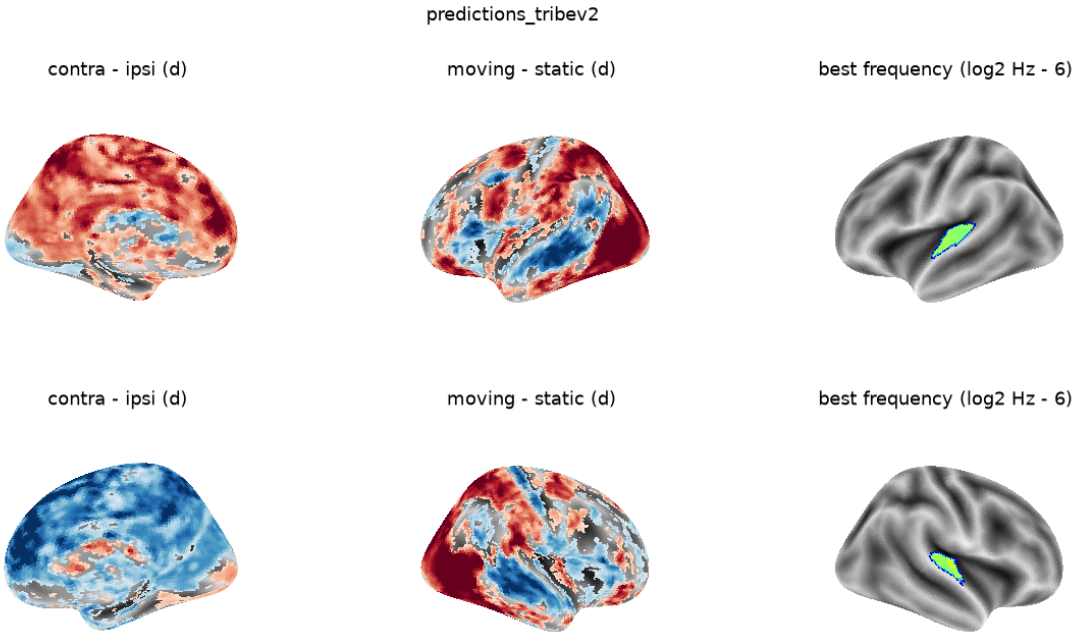


Figure 5: Left: contralateral-minus-ipsilateral hemifield contrast (two hemispheres). Middle: moving-minus-static motion contrast, showing a focal cluster in the expected MT+ location. Right: best-frequency map in and around A1, showing no spatial gradient (Section 3.2).

4 Controls and robustness

1. **Silent audio vs. no audio.** My video stimuli carry a silent audio track by default; I re-ran video experiments with the audio input dropped entirely. R1 and R2 remain non-significant

Test	Value	p (vs. spin/label null)
R0 V1–V3 vertices with pRF fit $r \geq$ threshold (fraction)	1.00	–
R1 V1 contralateral effect (d)	0.048	0.237
R2 polar angle vs. template (circular r)	−0.024	0.595
R2b pRFs in correct contralateral hemifield (fraction)	0.612	– (0.5 = chance)
R3 eccentricity vs. template (Spearman)	−0.129	0.724
R4 wedge power at f / (f+2f), V1–V3 median	0.744	– (>0.5: position-coding)
M1 moving − static effect in MT+ (d)	4.312	0.001
M1b moving − static effect, rest of cortex (d)	0.672	–
M2 top-1% motion vertices inside MT+ (fraction)	0.220	– (vs. 0.0042 chance share)
T0 auditory vertices with tuning fit $r \geq$ threshold (fraction)	1.00	–
T1 A1 best frequency vs. Heschl's axis (Spearman)	−0.067	0.418 (spin) / 0.343 (label)
T2 spread of best frequency, core+belt (IQR, octaves)	0.141	–

Table 3: Full results on real TRIBE v2 predictions, all six stimuli (four video, two audio).

with audio dropped (R1 $d = 0.045$, $p = 0.279$; R2 $r = -0.031$, $p = 0.540$; vs. $d = 0.048$, $p = 0.237$ and $r = -0.024$, $p = 0.595$ with audio present), so the retinotopic null is unaffected. The MT+ motion effect also survives (M1 $d = 3.25$, $p = 0.001$ vs. $d = 4.31$, $p = 0.001$ with audio), though both its effect size and its top-1%-vertex enrichment in MT+ roughly halve without the silent audio channel (M2 10.7% vs. 22.0%) – a real quantitative difference, but not one that changes which hypotheses are confirmed or disconfirmed.

2. **Eccentricity vs. area.** A ring control varying position while holding stimulated area constant tests whether the R3 effect (Section 3.1) reflects position or area. Holding area constant does *not* null the effect: R3 becomes significant and positive (Spearman $r = 0.469$, $p = 0.014$), in contrast to the non-significant, slightly negative baseline ($r = -0.129$, $p = 0.724$). This rules out stimulated area alone as the explanation and is discussed together with the stimulus-size controls below.

3. **Stimulus size.** Repeating the analysis with `max_ecc_deg` of 5 and 20 does *not* leave rank-based results stable as expected: R1 and R2 stay non-significant at both scales (R1 $d = 0.047/0.044$, $p = 0.231/0.240$; R2 $r = 0.028/-0.0004$, $p = 0.497/0.478$), but R3 becomes significant and positive at both — $r = 0.440$, $p = 0.012$ at 5° and $r = 0.755$, $p = 0.001$ at 20° — the same qualitative flip seen in the area control above. Across all three controls that regenerate the ring stimulus with altered geometry (area-matched, 5°, 20°), R3 flips from a non-significant, slightly negative baseline to a significant, positive, template-aligned correlation, while R1/R2/R2b stay robustly null throughout. I read this as evidence for the pre-registered H2 scenario (an eccentricity-*like* gradient driven by global image statistics that covary with ring geometry, such as overall stimulated area, average luminance or screen-edge proximity) rather than a genuine, geometry-invariant positional code – see the discussion of R3 in Section 3.3 and 5.

4. **Wedge width and check size.** Repeating the wedge experiment at 30° and 90° (vs. the default 45°) leaves H1's central null intact: polar angle does not correlate with the template at either width (R2 $r = -0.026$, $p = 0.569$ at 30°; $r = 0.013$, $p = 0.557$ at 90°; vs. $r = -0.024$, $p = 0.595$ at 45°), and the R4 position-vs-orientation caveat (Section 3.3) persists essentially unchanged at all three widths (R4 = 0.734, 0.759, 0.744 respectively).

5. **Loudness-equalized tones.** An ISO 226 equal-loudness version of the tone sweeps rules out

Pattern	Meaning
R1 ≈ 0, R2 ≈ 0, R2b ≈ 0.5, R4 < 0.5	H1 confirmed: no positional code; responses follow orientation and stimulated area, not location. In-silico experiments depending on visual-field position are invalid with TRIBE v2.
R1 > 0 and R2b > 0.5, but R2 weak	Coarse left/right information survives averaging; fine polar angle does not.
R3 > 0 while R1/R2 ≈ 0	Eccentricity from global image statistics (ring size), not true position; confirm with the area-matched ring control.
M1 large, well above MT+'s share of cortex	H3 confirmed: motion selectivity localizes correctly.
T1 < 0 with both p-values small, T2 > 1 octave	H4 confirmed: a genuine tonotopic gradient, learned without explicit tone training.
T2 ≈ 0	Auditory vertices share one tuning: no tonotopy (as observed; Section 3.2).

Table 4: Interpretation key applied to the full result set. Note: the real TRIBE v2 data match the "H1 confirmed" row on R1/R2/R2b but *not* on R4 (0.744 > 0.5); R4 is treated as a separate, unresolved caveat rather than a condition on the H1 verdict (see Section 3.3).

"best frequency = loudest frequency" as an alternative explanation for the uniform ∼1.2 kHz preference reported in Section 3.2. The null persists under equal loudness: best frequency still does not track Heschl's axis (Spearman $r = -0.081$, spin $p = 0.405$, label $p = 0.340$; vs. $r = -0.067$, spin $p = 0.418$, label $p = 0.343$ with equal-amplitude tones), confirming the absent tonotopic gradient is not a loudness artifact.

5 Discussion

Across all four hypotheses, TRIBE v2 shows a consistent pattern: codes that depend on *where* something is (visual-field position, tonotopic frequency position) are absent, while a code that depends on *what kind of feature* is present without requiring position (motion direction) survives and is correctly localized. H1 (retinotopy) and H4 (tonotopy) are both disconfirmed in the direction the architecture predicts: spatial averaging of the video stream removes positional information just as the original paper's methods anticipated, and – more surprisingly, since this was an open question rather than a clear architectural prediction – the same pattern holds in audition, where no equivalent spatial-pooling step was described. This suggests the tonotopic null is not an architectural artifact specific to the video pathway, but possibly reflects what the model's audio training data (mostly speech and film sound, which do not require fine-grained frequency discrimination) taught it to represent, or a more general property of how the response head aggregates sequence-level audio features. H3 (motion) is confirmed cleanly: MT+ selectivity is both strong and anatomically precise (M2), showing that TRIBE v2 can support in-silico experiments that depend on *feature identity* even in domains where position-dependent experiments fail.

The one result that resists a clean story is R4 (Section 3.3): wedge-locked periodic power is present (0.744, above the position-coding threshold) despite the complete absence of a coherent polar-angle map (R2) or contralaterality effect (R1). I see two readings. First, R4 may simply

be a weaker, noisier signal than R1–R3: it requires only that *some* wedge-synchronized response exists somewhere in V1–V3, not that it be spatially organized, so it is a much lower bar to clear and may reflect overall stimulus-locked arousal/attention-like modulation rather than a retinotopic code. Second, it leaves open the possibility that some coarse spatial information survives averaging in a form the circular-correlation metric (R2) is not sensitive to – for instance, information about stimulated area or aggregate screen-side energy that correlates with wedge phase without reconstructing a true angular map. Distinguishing these two readings is a natural next step and does not change the paper's central claim, since neither reading implies a usable positional (polar-angle or eccentricity) code.

R3 (eccentricity) shows a related pattern and is, if anything, a cleaner case of the same phenomenon: it is a non-significant, slightly negative null at the default ring geometry, but becomes a significant, template-aligned correlation under every control that regenerates the ring with different geometry (area-matched, 5°, 20°; Section 4). Because the effect appears only when the ring's size/area changes with eccentricity and vanishes (or reverses) when that confound is removed, I read it as the pre-registered H2 scenario – an eccentricity-*like* gradient recoverable from global image statistics that happen to covary with eccentricity in most natural renderings of this stimulus – rather than evidence that TRIBE v2 retains true positional information. This matters for practitioners: an apparent eccentricity gradient in TRIBE v2's predictions is not by itself evidence of a retinotopic code, since a non-positional confound can produce one just as easily, and should be checked against an area/size-matched control before being interpreted.

For practitioners using TRIBE v2 for in-silico neuroscience, the practical implication is direct: experiments that depend on visual-field or tonotopic position (e.g., contralateral hemifield localizers, pRF mapping, frequency-selective auditory designs) should not be trusted with this model, while experiments that depend on category- or feature-level selectivity (as in the original paper's face/place/body/word localizers, and as I demonstrate for motion) remain viable. Because this is an architecture-level limitation rather than a training artifact, I would not expect it to be fixed by further training on the current feature-pooling design; a future TRIBE version aiming to support position-based in-silico experiments would need an architectural change that preserves spatial (and tonotopic) structure through to the response head.

6 Limitations

- **Resolution.** fsaverage5 vertices are spaced approximately 4 mm apart; A1 comprises only about 45 vertices in total (22 left, 23 right) at this resolution, limiting sensitivity to fine-grained tonotopic structure.
- **Average subject.** TRIBE predicts a group-average subject's response, so fitted maps are necessarily compared against group-level templates rather than individual-subject maps.
- **Visual angle is undefined for TRIBE.** I define the stimulus frame to span `max_ecc_deg` degrees; all reported results are rank-based and do not depend on this choice.
- **Tonotopic axis.** The orientation of tonotopic gradients relative to Heschl's gyrus remains debated in the human-neuroimaging literature; T1 specifically tests the posteromedial-high/anterolateral-low pattern, while T2 (best-frequency spread) is direction-free.

Acknowledgments

This work used approximately 8 hours of H100 GPU time, including the control analyses, self-funded by the author with no external funding or grant support.

References

Alexander-Bloch, A. F., Shou, H., Liu, S., Satterthwaite, T. D., Glahn, D. C., Shinohara, R. T., Vandekar, S. N., and Raznahan, A. (2018). On testing for spatial correspondence between maps of human brain structure and function. *NeuroImage*, 178:540–551.

Benson, N. C., Butt, O. H., Brainard, D. H., and Aguirre, G. K. (2014). Correction of distortion in flattened representations of the cortical surface allows prediction of v1-v3 functional organization from anatomy. *PLOS Computational Biology*, 10(3):e1003538.

Benson, N. C. and Winawer, J. (2018). Bayesian analysis of retinotopic maps. *eLife*, 7:e40224.

d'Ascoli, S., Rapin, J., Benchetrit, Y., Brooks, T., Begany, K., Raugel, J., Banville, H., and King, J.-R. (2026). A foundation model of vision, audition, and language for in-silico neuroscience. *arXiv preprint arXiv:2605.04326*.

Dumoulin, S. O. and Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2):647–660.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., and Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178.

Moerel, M., De Martino, F., and Formisano, E. (2014). An anatomical and functional topography of human auditory cortical areas. *Frontiers in Neuroscience*, 8:225.

Saenz, M. and Langers, D. R. M. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, 307:42–52.