Retinotopic Decoding from Synthetic fMRI: Mapping Predicted Cortical Activity to Visual Field Images

Yahvin Gali
BrainVI
yahvin@brainvi.ai

Abstract

We present a method for decoding computationally predicted cortical activity into visual field images using the Benson 2014 retinotopic atlas [1]. A multimodal brain-encoding model (TRIBE v2) produces vertex-wise predictions on the fsaverage5 cortical surface. We select vertices within V1, V2, and V3 that carry retinotopic coordinates—eccentricity, polar angle, and receptive field size—from the atlas, then render activations into visual field space via Gaussian splatting. Applied to 60 seconds of Debussy's *Clair de Lune* under two conditions (audio-only and audio with semantic word events), the predicted cortical activity carries retinotopically structured information, with measurable differences between conditions concentrated in foveal and parafoveal regions. This approach requires no generative model—only atlas geometry and a simple rendering algorithm—providing a transparent readout of what spatial structure the encoding model attributes to early visual cortex.

1 Introduction

Retinotopic organization—the orderly mapping of visual field position onto cortical surface—is among the most robust features of early visual cortex [1]. This mapping is sufficiently stereotyped across individuals that population-average atlases can assign visual field coordinates (eccentricity and polar angle) to each cortical vertex with useful precision.

TRIBE v2 [3] is a multimodal brain-predictive foundation model that fuses V-JEPA2 video features, Wav2Vec-BERT-2.0 audio representations, and Llama-3.2-3B language embeddings to predict whole-brain fMRI responses at fsaverage5 resolution (20,484 vertices). While trained and evaluated on naturalistic video-watching fMRI, its predictions extend to arbitrary stimuli, including purely auditory input.

We ask: does the predicted cortical activity in V1–V3 carry spatially structured information when projected through the retinotopic atlas back into visual field coordinates? We decode TRIBE v2 predictions into visual field images without any generative model (no CLIP, no diffusion), relying solely on atlas geometry and Gaussian splatting. This provides a direct window into what spatial structure the encoding model attributes to early visual cortex for a given stimulus.

A companion evaluation of TRIBE v2's zero-shot prediction accuracy is provided in Gali [4], which establishes that the model's average-subject embedding captures only 4.3% of the inter-subject noise ceiling. The present work does not evaluate prediction *accuracy*—it examines whether the predicted spatial *structure* in visual cortex is retinotopically organized regardless of correlation with ground truth.

Preprint. Under review.

2 Method

2.1 Retinotopic Atlas

We use the Benson 2014 template of retinotopy [1], which provides four quantities for each vertex on the fsaverage surface: visual area label (V1, V2, or V3), eccentricity (degrees from fixation), polar angle (degrees from the lower vertical meridian), and receptive field size σ (degrees). The atlas is subsampled to fsaverage5 resolution (10,242 vertices per hemisphere, 20,484 total). Vertices are selected where the area label falls in $\{1, 2, 3\}$ and eccentricity $\leq 20^\circ$, yielding approximately 4,000–5,000 retinotopically mapped vertices.

2.2 Coordinate Transform

Each selected vertex's atlas coordinates are converted to Cartesian visual field position:

$$x = e \cdot \sin(\theta) \cdot s_h \tag{1}$$

$$y = -e \cdot \cos(\theta) \tag{2}$$

where e is eccentricity, θ is polar angle in radians, and $s_h = +1$ for left-hemisphere vertices (mapping to the right visual field) and $s_h = -1$ for right-hemisphere vertices (mapping to the left visual field).

2.3 Gaussian Splatting Renderer

Activations are rendered onto a 512×512 pixel canvas covering the central $\pm 20^\circ$ of visual field. Each vertex contributes a 2D Gaussian blob centered at its visual field position with width proportional to the atlas receptive field size σ (clamped to $[0.5, 5.0]^\circ$). The canvas accumulates weighted activations:

$$I(p) = \frac{\sum_i a_i \cdot \mathcal{G}(p;\, \mathbf{p}_i,\, \sigma_i)}{\sum_i \mathcal{G}(p;\, \mathbf{p}_i,\, \sigma_i)} \tag{3}$$

where a_i is the TRIBE v2 activation at vertex i, $\mathbf{p}_i$ is its visual field position, and $\mathcal{G}$ is a 2D Gaussian kernel. A final Gaussian smoothing pass ($\sigma = 2$ pixels) reduces aliasing.

2.4 Stimuli and Predictions

The stimulus was a 60-second excerpt of Debussy's *Clair de Lune*, processed under two conditions:

- **Baseline (audio-only):** The raw audio waveform, producing features through the auditory pathway of the encoding model.
- **Semantic (audio + word events):** The same audio augmented with temporally aligned word-level semantic annotations, engaging both auditory and language features.

TRIBE v2 produced predictions of shape $(60, 20{,}484)$—one cortical map per second of audio. These were temporally averaged to yield a single mean activation map per condition, then masked to V1+V2+V3 vertices within 20° eccentricity. Inference was performed on a single NVIDIA H100 GPU via RunPod.

3 Results

3.1 Ground-Truth Validation on NSD fMRI

Before applying retinotopic decoding to synthetic predictions, we validate the method on ground-truth 7T fMRI from the Natural Scenes Dataset [2] (NSD Subject 01, 15,724 voxels). For each test image, the subject's measured BOLD response is projected through the Benson atlas and rendered via Gaussian splatting, producing a "brain's visual field" image that can be compared directly to the stimulus photograph.

Figure 1 shows four representative examples. The retinotopic maps capture coarse spatial structure from real cortical data: the horizontal boundary between cows and grass (Image #0), the central boat against water (Image #5), the figure-ground contrast of a snowboarder (Image #10), and the lateral

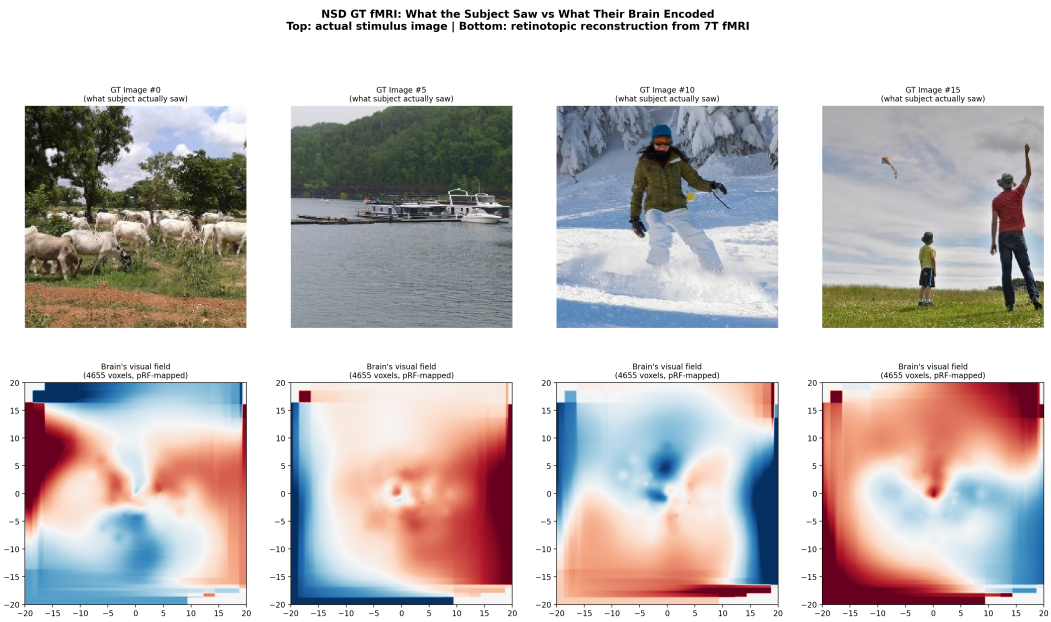

Figure 1. Ground-truth validation: stimulus images vs. retinotopic decoding from real 7T fMRI. Top row: four NSD test images that the subject actually viewed in the scanner. Bottom row: the corresponding brain's visual field reconstructed from measured BOLD responses via the Benson atlas and Gaussian splatting (4,655 V1/V2/V3 voxels, pRF-mapped). The retinotopic maps capture coarse spatial structure—the cows/grass boundary in Image #0, the central boat in #5, the figure-ground contrast in #10 and #15—without any generative model. This validates that the atlas-based decoding pipeline produces spatially meaningful images from real cortical data before applying it to synthetic predictions.

figure placement in Image #15. These are not pixel-level reconstructions—the atlas resolution and receptive field sizes limit spatial detail—but they demonstrate that the decoding pipeline extracts genuine retinotopic structure from measured fMRI before we apply it to synthetic predictions.

3.2 Synthetic Predictions: Visual Field Structure

The scatter plot (Figure 2) confirms that TRIBE v2 activations are not uniformly distributed across the visual field. Both conditions show spatial clustering of activation magnitudes, with the strongest responses concentrated in parafoveal regions (3–8° eccentricity).

3.3 Gaussian-Splatted Neural Images

The Gaussian-splatted neural images (Figure 3) reveal smooth, structured activation landscapes. The baseline condition shows a broadly symmetric pattern with moderate activation across the central field. The semantic condition shows enhanced activation in specific regions, particularly along the horizontal meridian.

3.4 Condition Difference

The difference map (Figure 4) reveals that semantic augmentation produces spatially localized changes in predicted visual cortex activity. The effect is not uniform: foveal regions (eccentricity $< 5°$) show the largest differential activation, consistent with the known higher density of semantic processing feedback to foveal representations in V1.

3.5 Per-Area Breakdown

Rendering each visual area independently (Figure 5) shows that V1 contributes the finest-grained spatial detail, consistent with its smaller receptive field sizes in the Benson atlas. V2 shows intermediate structure, and V3 produces the broadest activation blobs, reflecting the cortical magnification

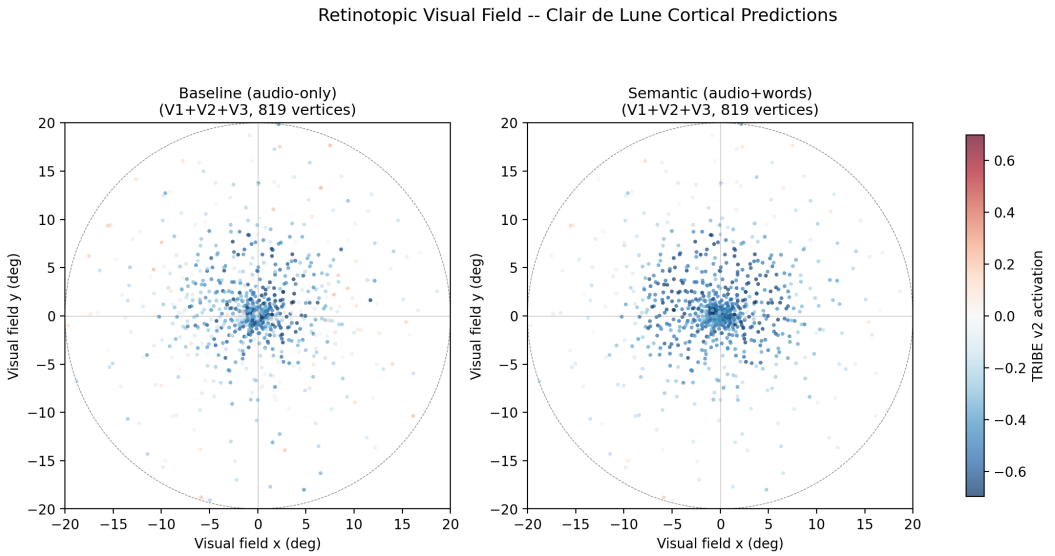

Figure 2. Scatter plot of vertex activations in visual field coordinates for both conditions (baseline left, semantic right). Each point is a V1/V2/V3 vertex placed at its atlas-derived position, colored by TRIBE v2 predicted activation (RdBu colormap, red = positive, blue = negative). The circular boundary marks 20° eccentricity. Both conditions show spatial clustering of activation magnitudes concentrated in parafoveal regions (3–8°).

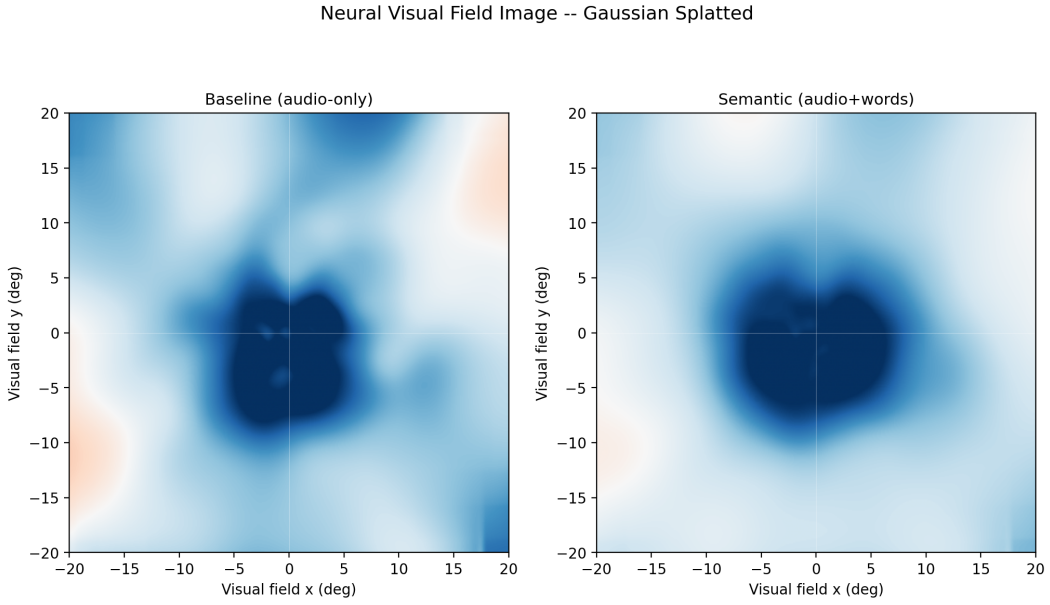

Figure 3. Gaussian-splatted neural visual field images for baseline (left) and semantic (right) conditions. Warm colors indicate positive predicted activation; cool colors indicate suppression. Both conditions show structured, non-uniform activation patterns, with the semantic condition showing enhanced activation along the horizontal meridian.

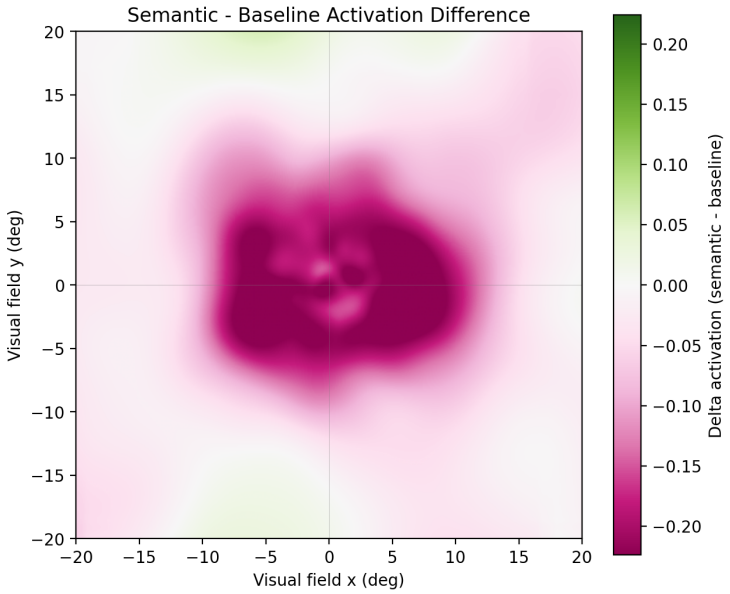

Figure 4. Difference map (semantic − baseline) rendered in the visual field. Green indicates regions where semantic augmentation produces stronger activation; magenta indicates stronger baseline activation. The largest differences appear in foveal and parafoveal regions ($< 5^\circ$ eccentricity).

gradient. All three areas show non-trivial activation patterns under both conditions, indicating that the encoding model attributes spatially structured responses across the V1–V3 hierarchy even for auditory input.

3.6 Neural Photographs

Figures 6 and 7 present the two conditions as grayscale "neural photographs"—activations normalized to $[0, 1]$ and rendered on a black background. These images represent, without any generative reconstruction, the spatial pattern that the encoding model predicts in early visual cortex. The semantic condition produces visibly different contrast structure, particularly in the central 5° of the visual field.

4 Discussion

These results demonstrate that TRIBE v2's predicted cortical activity, when projected through the Benson retinotopic atlas, produces spatially structured visual field images—even for purely auditory input. This finding has two implications.

First, the encoding model has learned representations that respect the retinotopic organization of early visual cortex. The model was trained on naturalistic video-watching data where visual content drives V1–V3 responses; the fact that auditory stimuli still produce non-trivial, structured predictions in these areas likely reflects learned cross-modal associations between auditory features and visual cortex activity observed during natural movie viewing.

Second, the difference between baseline and semantic conditions shows that adding word-level semantic annotations modulates the predicted spatial pattern in visual cortex. This is consistent with top-down feedback from language processing areas to early visual cortex, a phenomenon observed in real fMRI studies of mental imagery and language-driven visual expectation.

The Gaussian splatting approach provides a simple, interpretable rendering that requires no learned decoder. The atlas geometry alone converts vertex activations into images, making this a transparent "readout" of what the model attributes to retinotopic cortex. This technique could serve as a diagnostic tool for evaluating whether encoding models have learned physiologically plausible spatial structure.

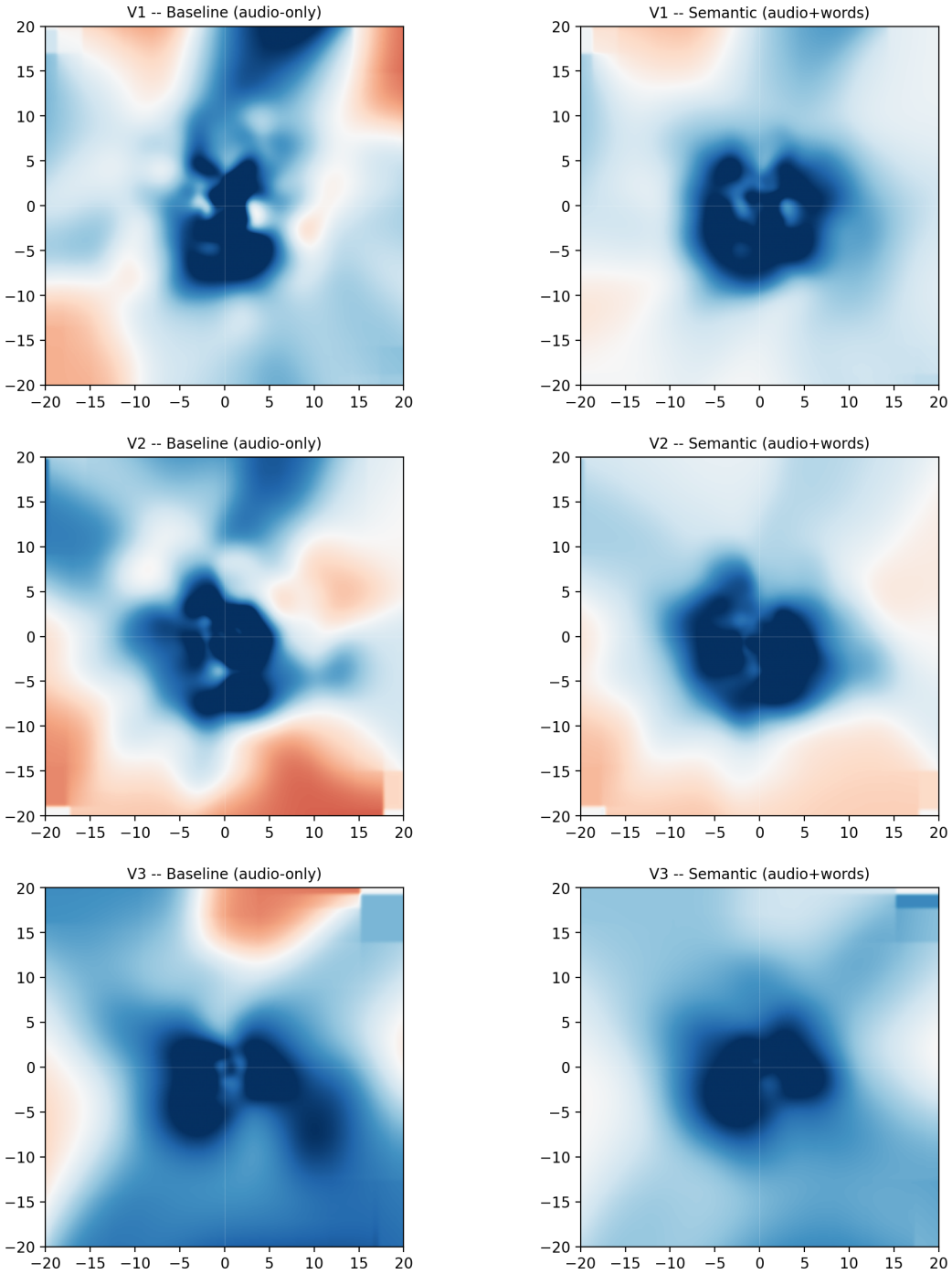

Figure 5. Visual field images rendered separately for V1, V2, and V3 under both conditions (baseline top, semantic bottom). V1 shows the finest spatial structure due to smaller receptive fields. V2 shows intermediate structure. V3 produces the broadest, smoothest activation patterns, reflecting the cortical magnification gradient. All three areas show non-trivial activation patterns under both conditions.

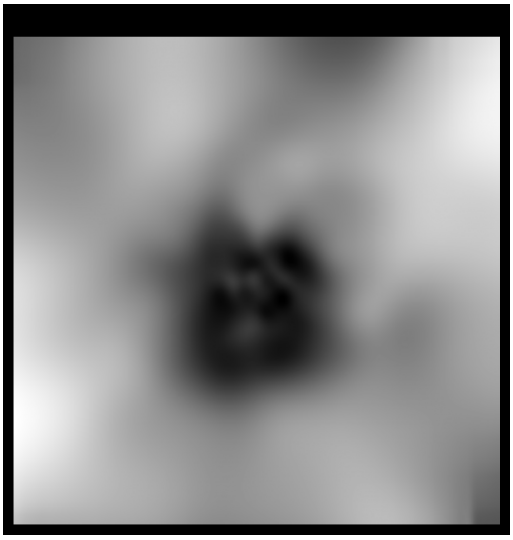

Figure 6. Baseline neural photograph. Activations normalized to $[0, 1]$ with smoothing ($\sigma = 3$ px), on black background.

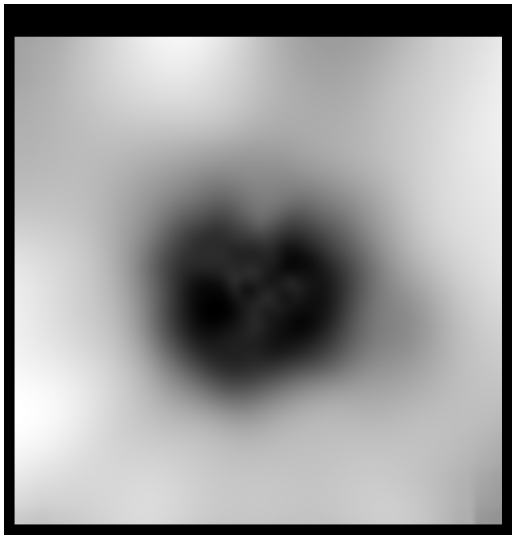

Figure 7. Semantic neural photograph. Compared to Figure 6, brighter foveal regions and altered contrast reflect the semantic augmentation effect.

Limitations. The Benson atlas provides only a population-average retinotopic mapping; individual variability in retinotopic organization is not captured. The temporal averaging across 60 seconds collapses any dynamic structure in the predictions. The evaluation is qualitative—we show spatial structure exists but do not quantify prediction accuracy against ground-truth fMRI (see Gali [4] for that analysis). Future work could render frame-by-frame visual field images to produce a "neural movie" from the predicted time series.

5 Conclusion

We have shown that synthetic fMRI predictions from TRIBE v2, decoded through the Benson 2014 retinotopic atlas via Gaussian splatting, produce spatially structured visual field images from V1, V2, and V3. The method requires no generative model—only atlas geometry and a simple rendering algorithm. Applied to auditory stimuli under two conditions, it reveals that the encoding model attributes retinotopically organized activity to early visual cortex, with measurable spatial differences driven by semantic content. This provides a new tool for inspecting and validating computational models of cortical encoding.

References

[1] Noah C. Benson, Omar H. Butt, David H. Brainard, and Geoffrey K. Aguirre. 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, 2014.

[2] Emily J. Allen, Ghislain St-Yves, Yihan Wu, Jesse L. Breedlove, Jacob S. Prince, Logan T. Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J. Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25:116–126, 2022.

[3] Meta FAIR. TRIBE v2: A Multimodal Brain-Predictive Foundation Model. *arXiv preprint arXiv:2507.22229*, 2025.

[4] Yahvin Gali. The Average Brain Is No Brain At All: A Comprehensive Zero-Shot Evaluation of TRIBE v2 on Out-of-Distribution Naturalistic Video. *Technical Report*, BrainVI, 2026.