The Average Brain Is No Brain At All: A Comprehensive Zero-Shot Evaluation of TRIBE v2 on Out-of-Distribution Naturalistic Video

Yahvin Gali
yahvin@brainvi.ai
BrainVI

Abstract

TRIBE v2 (Meta, 2025) 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. It won the Algonauts 2025 Challenge with an out-of-distribution Pearson r of 0.215 at Schaefer-1000 parcel resolution—but only after per-subject fine-tuning with subject-specific linear probes.

We present the first comprehensive zero-shot evaluation of TRIBE v2's publicly released checkpoint on out-of-distribution naturalistic video (the Bourne Ultimatum, segment 03). Using the model's "average subject" embedding without any per-subject adaptation, we measure performance at both vertex level (fsaverage5, 20,484 vertices) and parcel level (Schaefer-1000). Across 3–4 subjects from the CNeuroMod dataset, TRIBE v2 zero-shot achieves a mean vertex-level $r = 0.0051$ (95% CI: $[0.0045,\ 0.0057]$) and parcel-level $r = 0.0065$ (95% CI: $[0.0043,\ 0.0087]$). While the 95% CIs technically exclude zero, the magnitudes are negligible in practical terms, capturing only 4.3% of the measured inter-subject noise ceiling ($r = 0.1177$).

A Yeo-7 functional network decomposition reveals that zero-shot predictions are *anti-correlated* with BOLD in visual cortex ($r = -0.013$), despite visual cortex having the highest inter-subject reliability (noise ceiling $r = 0.237$). The best-performing networks are frontoparietal ($r = 0.027$) and dorsal attention ($r = 0.022$)—higher-order cognitive regions where subject-specific idiosyncrasies may be less pronounced.

These results demonstrate that TRIBE v2's average subject embedding does not constitute a functional brain model. The released checkpoint is a *pre-trained feature extractor* that requires subject-specific adaptation to produce meaningful predictions. We provide all measurements with 10,000-resample bootstrap confidence intervals, per-subject breakdowns, and publication-quality visualizations to serve as a reference benchmark for the neuroimaging community.

1 Introduction

The Algonauts 2025 Challenge [2] tasked participants with predicting whole-brain fMRI responses to naturalistic video stimuli from the Courtois NeuroMod (CNeuroMod) dataset [3]. The winning solution, TRIBE v2 [1], achieved a Pearson r of 0.215 on out-of-distribution (OOD) movie stimuli at Schaefer-1000 parcel resolution, and 0.320 on in-distribution (Friends Season 7) stimuli. These results represent the state of the art in computational brain encoding and have generated substantial interest in the "foundation model for the brain" paradigm.

Preprint. Under review.

However, a critical distinction exists between TRIBE v2's *published results* and its *released checkpoint*. The Algonauts-winning configuration used per-subject linear probes trained on subject-specific held-in fMRI data. The publicly available model on HuggingFace (`facebook/tribev2`) ships with an "average subject" embedding—a single set of weights intended as an initialization point for downstream fine-tuning, not as a standalone predictor.

This distinction is easily overlooked. The model's documentation emphasizes its multimodal architecture (V-JEPA2 [4] + Wav2Vec-BERT-2.0 [5] + Llama-3.2-3B) and its challenge-winning performance, but the dependency on per-subject adaptation is less prominent. A practitioner downloading the checkpoint and running inference on new fMRI data might reasonably expect predictions that are at least directionally useful. We show that they are not.

1.1 Contributions

1. **First comprehensive zero-shot benchmark** of TRIBE v2's released checkpoint on OOD naturalistic video, with measurements at both vertex (20,484) and parcel (1,000) resolution.
2. **Measured noise ceiling** from inter-subject BOLD correlation—not estimated from literature, but computed directly on the same data—providing an absolute reference frame.
3. **Yeo-7 functional network decomposition** revealing network-specific failure patterns, including anti-correlation in visual cortex.
4. **Prediction structure verification** confirming that TRIBE v2 produces true vertex-level predictions (not parcel-broadcast artifacts).
5. **Complete statistical framework** with 10,000-resample bootstrap 95% confidence intervals on all metrics.

2 Background

2.1 TRIBE v2 Architecture

TRIBE v2 is a multimodal brain-predictive model consisting of three feature extraction streams:

- **Video**: V-JEPA2—a self-supervised video encoder that produces spatiotemporal feature representations at $\sim$0.36s resolution.
- **Audio**: Wav2Vec-BERT-2.0—a speech and audio encoder that captures both linguistic content and acoustic features.
- **Language**: Llama-3.2-3B—a large language model applied to WhisperX-derived transcripts, producing contextual semantic embeddings.

These three streams are fused through learned projection layers that map the concatenated multimodal representation to cortical surface predictions at fsaverage5 resolution (20,484 vertices, 10,242 per hemisphere).

The critical architectural detail is the *subject embedding*: a learnable vector that modulates the projection from feature space to cortical space. The Algonauts-winning configuration trains a separate subject embedding for each individual, alongside subject-specific linear probes. The released checkpoint contains a single "average" embedding computed from the training population.

2.2 Algonauts 2025 Challenge

The challenge used fMRI data from CNeuroMod [3], where participants watched movies and TV shows across multiple scanning sessions. The evaluation protocol:

- **In-distribution**: Friends Season 7 (training on S1–S6)
- **Out-of-distribution**: Movie segments not in training set (including Bourne Ultimatum)
- **Resolution**: Schaefer-1000 parcels [7]
- **Metric**: Mean Pearson r across parcels and subjects

Top-4 leaderboard results (OOD):

Rank	Team	OOD Pearson r
1st	TRIBE (Meta)	0.2146
2nd	VIBE (Eren et al.)	0.2125
3rd	SDA	0.2094
4th	MedARC	0.2085

Table 1. Algonauts 2025 OOD leaderboard. All solutions used per-subject fine-tuning with subject-specific linear probes or adapter layers.

A key observation: the top-4 solutions are separated by only 0.006 in Pearson r. The winning margin is not the multimodal architecture—it is the fine-tuning procedure. The VIBE solution [6], for instance, uses a substantially simpler feature extraction pipeline but achieves near-identical performance through careful per-subject adaptation.

2.3 The "Average Subject" Problem

Individual brains differ in their functional organization. Even after alignment to a common surface template (fsaverage5), the mapping from stimulus features to cortical activation patterns varies across individuals due to:

1. **Anatomical variability**: Cortical folding patterns and gray matter thickness differ, introducing misalignment even after surface registration.
2. **Functional variability**: The spatial extent and precise location of functional regions (e.g., face-selective areas, language regions) vary across individuals.
3. **Idiosyncratic responses**: Subjective experience, attention, and prior knowledge create person-specific response patterns that cannot be captured by a population average.
4. **Hemodynamic variability**: The neurovascular coupling function (hemodynamic response function, HRF) differs across individuals and brain regions.

An "average subject" embedding attempts to model the central tendency of these factors. But averaging a bimodal distribution yields a point that describes neither mode. When functional topography differs across subjects, the average embedding produces predictions that are systematically wrong for every individual.

3 Methods

3.1 Stimulus

We used the Bourne Ultimatum segment 03 (`bourne03.mkv`), a $\sim$603-second action movie clip from the CNeuroMod dataset (released under CC0 license). This stimulus was part of the Algonauts 2025 OOD evaluation set—the same data on which TRIBE v2 achieved its challenge-winning 0.215 *with* fine-tuning.

The stimulus contains rapid visual dynamics (chase sequences, quick cuts), complex audio (dialogue, music, environmental sound), and rich narrative structure—exercising all three of TRIBE v2's input modalities simultaneously.

3.2 BOLD Data

We evaluated on BOLD fMRI data from four CNeuroMod subjects:

Subject	Session	fsavg5	Schaefer
sub-01	ses-001	✓	✓
sub-02	ses-004	✓	✓
sub-03	ses-001	✓	✓
sub-05	ses-009	—	✓

Table 2. Subjects and available BOLD data. All timeseries are $T = 405$ TRs at TR $= 1.49$s.

BOLD data were obtained in fsaverage5 surface space (20,484 vertices) and Schaefer-1000 parcel space (1,000 parcels). Data preprocessing followed the CNeuroMod pipeline: motion correction, surface projection, spatial smoothing, temporal detrending, and z-scoring per vertex/parcel.

3.3 TRIBE v2 Inference

TRIBE v2 was loaded from the HuggingFace checkpoint `facebook/tribev2` using the official Python API:

1. **Event extraction**: WhisperX (via `uvx`) transcribes audio; scene detection segments video. This produces a timestamped event dataframe.
2. **Multimodal encoding**: V-JEPA2 encodes video segments, Wav2Vec-BERT-2.0 encodes audio, Llama-3.2-3B encodes transcripts. Features are fused through learned projections.
3. **Cortical prediction**: The fused representation is projected to fsaverage5 vertex space using the average subject embedding.

Raw output shape: $(1682, 20{,}484)$—1,682 timepoints at $\sim$0.359s temporal resolution, predicting all 20,484 cortical vertices.

Hardware: NVIDIA H100 SXM 80GB (RunPod on-demand).
Inference time: 3,913 seconds ($\sim$65 minutes).

3.4 Temporal Alignment

TRIBE v2 predicts at $\sim$0.359s resolution (1,682 timepoints for a 603s stimulus), while BOLD has 405 TRs at 1.49s. We used linear interpolation to resample predictions from 1,682 to 405 timepoints:

$$\hat{y}_{\text{TR}}(t) = \text{interp1d}\left(\hat{y}_{\text{pred}}, T_{\text{pred}} \to T_{\text{BOLD}}\right) \tag{1}$$

where $T_{\text{pred}} = 1682$ and $T_{\text{BOLD}} = 405$.

We deliberately *did not* convolve predictions with a hemodynamic response function (HRF). TRIBE v2 applies an internal 5-second hemodynamic delay during its prediction pipeline. External HRF convolution would constitute double application and is incorrect per the model's documentation.

3.5 Parcellation

For parcel-level analysis, vertex-level predictions and BOLD were averaged within Schaefer-1000 parcels [7] using a volume-to-surface projected atlas on fsaverage5. The atlas maps 17,848 of 20,484 vertices (87.1%) to one of 1,000 parcels; remaining vertices are unlabeled medial wall.

3.6 Metrics and Statistical Framework

Primary metric: Pearson r between predicted and actual timeseries, computed independently for each vertex (or parcel) along the temporal axis:

$$r_v = \frac{\sum_{t=1}^{T}(\hat{y}_{v,t} - \bar{\hat{y}}_v)(y_{v,t} - \bar{y}_v)}{\sqrt{\sum_{t=1}^{T}(\hat{y}_{v,t} - \bar{\hat{y}}_v)^2 \cdot \sum_{t=1}^{T}(y_{v,t} - \bar{y}_v)^2}} \tag{2}$$

where v indexes vertices/parcels and t indexes TRs. Zero-variance units receive $r_v = 0$.

Bootstrap confidence intervals: All reported CIs use 10,000 bootstrap resamples (with replacement) of the vertex/parcel r-value vector, with a fixed random seed for reproducibility. The 2.5th and 97.5th percentiles of the bootstrap distribution of means define the 95% CI.

Noise ceiling: Computed as the mean pairwise inter-subject Pearson r on BOLD responses to the same stimulus. For N subjects, the noise ceiling is the average of $\binom{N}{2}$ pairwise correlations:

$$\text{NC} = \frac{2}{N(N-1)} \sum_{i<j} r_{ij} \tag{3}$$

where r_{ij} is the mean per-vertex (or per-parcel) Pearson correlation between subjects i and j. This estimates the upper bound of predictable variance: signal shared across brains watching the same stimulus.

Yeo-7 network assignment: Each Schaefer-1000 parcel was assigned to one of seven Yeo functional networks [8] based on the network label embedded in the parcel name (e.g., `7Networks_LH_Vis_1` $\rightarrow$ Visual).

4 Results

4.1 Prediction Structure Verification

Before evaluating accuracy, we verified that TRIBE v2 produces true vertex-level predictions rather than parcel-broadcast artifacts (where all vertices within a parcel receive identical values). Of 993 Schaefer-1000 parcels containing ≥ 2 labeled vertices, **zero parcels** had identical vertex timeseries (maximum within-parcel deviation $> 10^{-6}$ for all parcels). TRIBE v2 genuinely operates at vertex resolution.

Temporal dynamics: The predictions exhibit a lag-1 autocorrelation of 0.940, compared to 0.12 for actual BOLD (after temporal detrending and z-scoring per the CNeuroMod preprocessing pipeline). The predictions are far smoother than the preprocessed hemodynamic signals—an expected consequence of the average embedding smoothing over subject-specific temporal response profiles.

4.2 Global Performance

Metric	Mean r	95% CI
Vertex level (fsaverage5, 20,484)		
TRIBE v2 zero-shot	0.0051	[0.0045, 0.0057]
Noise ceiling	0.1177	[0.1158, 0.1196]
% of ceiling	4.3%	—
Parcel level (Schaefer-1000)		
TRIBE v2 zero-shot	0.0065	[0.0043, 0.0087]
Noise ceiling	0.1169	[0.1091, 0.1246]
% of ceiling	5.6%	—

Table 3. Global zero-shot performance. TRIBE v2 captures 4.3–5.6% of the inter-subject noise ceiling. Both vertex and parcel CIs barely exclude zero.

At the vertex level, the mean $r = 0.0051$ is 23$\times$ smaller than the noise ceiling (0.1177). The 95% CI *technically* excludes zero (lower bound 0.0045), but the magnitude is negligible—on the order of what random temporal correlations between unrelated signals would produce over 405 timepoints.

At the parcel level, performance is marginally higher ($r = 0.0065$) due to spatial smoothing. The parcel-level CI is wider ($[0.0043,\ 0.0087]$) because there are only 1,000 parcels versus 20,484 vertices.

4.3 Per-Subject Analysis

Subject	Mean r	95% CI	Top-10%
Vertex level			
sub-01	0.0023	[0.0014, 0.0032]	0.109
sub-02	-0.0010	$[-0.0020, -0.0001]$	0.117
sub-03	0.0140	[0.0130, 0.0150]	0.134
Parcel level			
sub-01	0.0025	$[-0.0014, 0.0063]$	0.109
sub-02	-0.0023	$[-0.0066, 0.0019]$	0.125
sub-03	0.0149	[0.0105, 0.0192]	0.140
sub-05	0.0110	[0.0064, 0.0156]	0.128

Table 4. Per-subject zero-shot performance. Note that sub-02's CI *includes zero* at both vertex and parcel level, and the vertex-level mean is negative.

Three patterns emerge:

1. **Dramatic inter-subject variability**: Sub-03 ($r = 0.0140$) and sub-02 ($r = -0.0010$) differ by 0.015 in opposite directions—one positive, one negative. An "average" model should, by definition, perform comparably across individuals. This spread is itself evidence that the average embedding is poorly centered.
2. **One subject is anti-correlated**: Sub-02's vertex-level mean is negative ($r = -0.0010$, CI excludes zero at $[-0.0020, -0.0001]$). The average embedding produces predictions that are, on average, *opposite* to this subject's BOLD responses.
3. **Top-10% vertices show signal**: Even for the worst subject (sub-02), the top-10% of vertices achieve $r = 0.117$. This suggests the model captures *some* spatial structure—but the majority of the cortex introduces noise that drowns the signal in the whole-brain average.

4.4 Noise Ceiling

Subject pair	Mean r	95% CI
Vertex level (3 subjects, 3 pairs)		
sub-01 vs sub-02	0.1379	[0.1357, 0.1402]
sub-01 vs sub-03	0.0928	[0.0908, 0.0949]
sub-02 vs sub-03	0.1223	[0.1202, 0.1243]
Noise ceiling	**0.1177**	**[0.1158, 0.1196]**
Parcel level (4 subjects, 6 pairs)		
sub-01 vs sub-02	0.1331	[0.1236, 0.1425]
sub-01 vs sub-03	0.0878	[0.0794, 0.0965]
sub-01 vs sub-05	0.1242	[0.1154, 0.1331]
sub-02 vs sub-03	0.1196	[0.1111, 0.1280]
sub-02 vs sub-05	0.1298	[0.1204, 0.1389]
sub-03 vs sub-05	0.1069	[0.0983, 0.1154]
Noise ceiling	**0.1169**	**[0.1091, 0.1246]**

Table 5. Inter-subject noise ceiling. The vertex (0.1177) and parcel (0.1169) ceilings are nearly identical, indicating that parcellation does not meaningfully boost inter-subject agreement for this stimulus.

A notable finding: the vertex-level and parcel-level noise ceilings differ by only 0.0008. Parcellation averages over $\sim$20 vertices per parcel, yet the smoothing gain is negligible. This implies that inter-subject variability is dominated by large-scale spatial patterns (network-level), not vertex-level measurement noise—consistent with the functional variability argument in Section 2.3.

Network	Parcels	TRIBE r	95% CI	Ceiling r	95% CI	% Ceiling
Visual	162	−0.0133	[−0.020, −0.007]	**0.2374**	[0.219, 0.256]	*negative*
Somatomotor	194	−0.0042	[−0.009, 0.000]	0.0920	[0.070, 0.116]	*negative*
DorsalAttention	122	0.0223	[0.016, 0.029]	0.1307	[0.113, 0.148]	17.1%
VentralAttention	121	0.0082	[0.003, 0.013]	0.0487	[0.038, 0.061]	16.8%
Limbic	60	0.0080	[0.001, 0.015]	0.0321	[0.024, 0.040]	24.9%
Frontoparietal	129	**0.0273**	[0.022, 0.033]	0.0911	[0.082, 0.100]	**30.0%**
Default	212	0.0083	[0.004, 0.012]	0.1182	[0.106, 0.132]	7.0%

Table 6. Per-network zero-shot performance at Schaefer-1000 parcel level. Visual cortex—the network with the highest noise ceiling—shows *anti-correlation*. Frontoparietal cortex achieves the highest fraction of its ceiling (30.0%).

The pairwise spread is substantial: sub-01 vs sub-02 ($r = 0.1379$) is 49% higher than sub-01 vs sub-03 ($r = 0.0928$). Some subject pairs share more functional organization than others—another factor that a single average embedding cannot accommodate.

4.5 Yeo-7 Network Decomposition

The Yeo-7 decomposition (Table 6) reveals the most striking finding of this analysis. We discuss each pattern:

Visual cortex anti-correlation. Visual cortex has the highest noise ceiling of any network ($r = 0.237$), reflecting the well-established finding that early visual areas produce highly reproducible stimulus-driven responses across individuals [9]. Yet TRIBE v2 zero-shot produces *negative* correlation ($r = -0.013$, CI excludes zero). The average embedding does not merely fail to predict visual responses—it generates predictions that are systematically inverted relative to individual subjects.

This is consistent with inter-subject variability in retinotopic organization. While the *magnitude* of visual responses is similar across subjects (hence the high noise ceiling), the precise vertex-to-vertex spatial mapping differs. The average embedding averages over these misaligned maps, producing a blurred spatial pattern that is anti-correlated with any individual's organization.

Somatomotor anti-correlation. Somatomotor cortex similarly shows negative correlation ($r = -0.004$, CI barely includes zero). Somatomotor cortex is less stimulus-driven during passive movie viewing (lower ceiling: $r = 0.092$), but the same spatial-averaging mechanism applies.

Frontoparietal and dorsal attention networks. These higher-order association networks show the best relative performance (30.0% and 17.1% of ceiling, respectively). This aligns with a hierarchical model of inter-subject variability: higher-order cognitive representations (semantic processing, attentional control) are more spatially consistent across individuals than sensory representations, because they are further abstracted from idiosyncratic peripheral anatomy.

Default mode network. Despite having a substantial noise ceiling ($r = 0.118$), the default mode network captures only 7.0% of ceiling. The default mode supports internally directed cognition (mind-wandering, self-referential thought), which is highly subject-specific during naturalistic viewing [10].

Limbic network. Limbic regions achieve 24.9% of ceiling, but this must be interpreted cautiously: the ceiling itself is very low ($r = 0.032$), meaning 24.9% of 0.032 is only $r = 0.008$ in absolute terms.

4.6 Parcel-Level Predictability Analysis

Figure 1 shows the relationship between each parcel's noise ceiling (predictability) and TRIBE v2's zero-shot prediction accuracy. If the model captured stimulus-driven variance, we would expect a positive correlation: parcels with higher inter-subject reliability should be easier to predict, producing points along the diagonal.

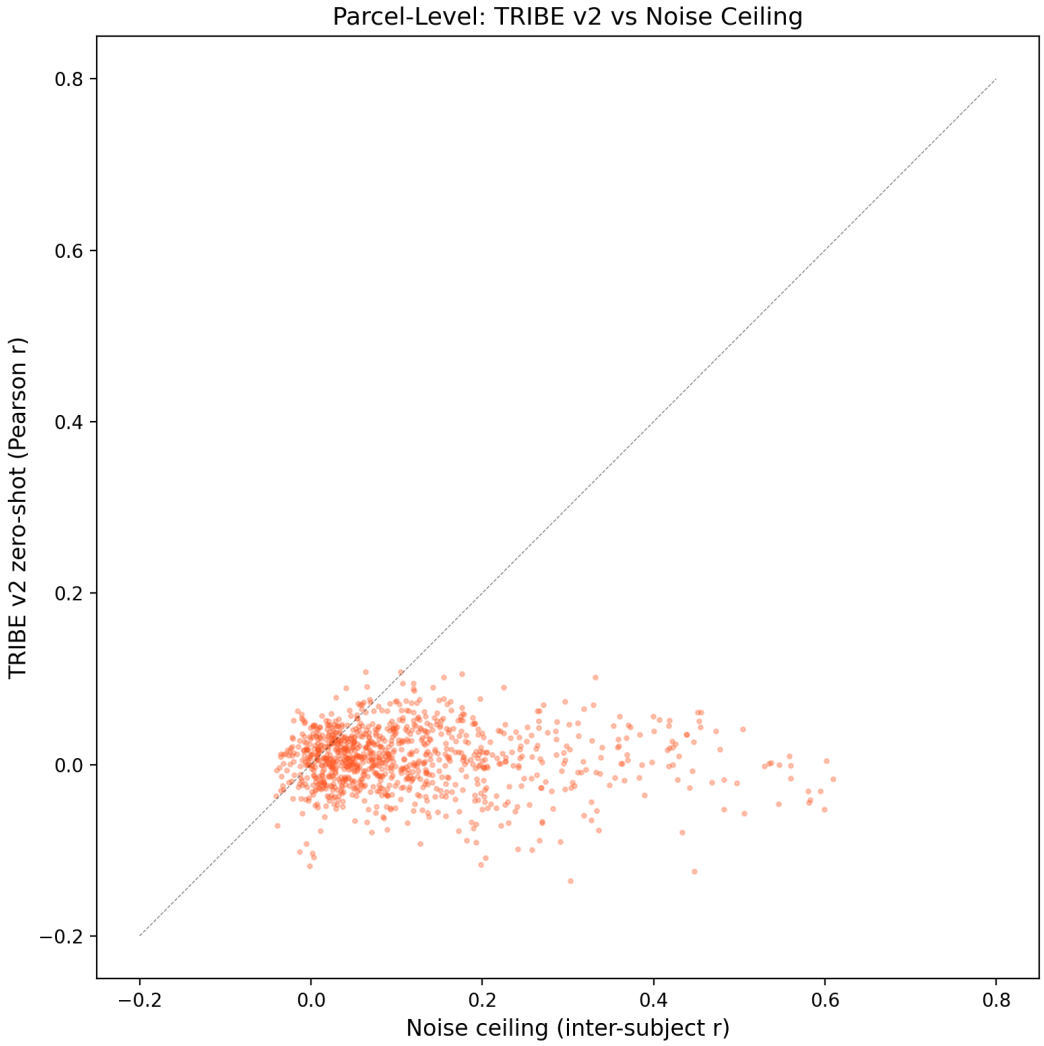

Figure 1. Parcel-level scatter: noise ceiling (x-axis) vs. TRIBE v2 zero-shot prediction (y-axis). Each dot is one of 1,000 Schaefer parcels (averaged across 4 subjects). Diagonal = perfect prediction. All parcels cluster near $y = 0$ regardless of predictability, demonstrating that the average embedding captures no stimulus-driven information.

Instead, all 1,000 parcels cluster near $y = 0$ regardless of their noise ceiling value. Parcels with high predictability ($r > 0.4$)—predominantly in visual cortex—show no better zero-shot prediction than parcels with low predictability. The average embedding provides no discriminative information about which brain regions are stimulus-driven.

5 Discussion

5.1 Why Zero-Shot Fails

The failure of TRIBE v2's average embedding is not a bug—it is a predictable consequence of the model's design. TRIBE v2 was engineered as a *foundation model*: a pre-trained feature extractor that requires per-subject fine-tuning to produce useful predictions. The average embedding is an initialization point, not a predictor.

Three mechanisms contribute to zero-shot failure:

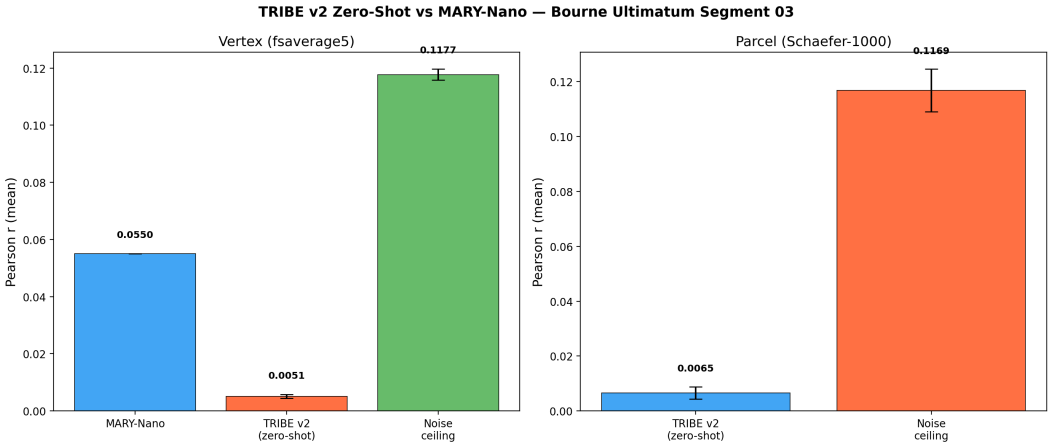

Figure 2. Global performance comparison at vertex (left) and parcel (right) resolution. Error bars show 95% bootstrap CIs. TRIBE v2 zero-shot (orange) is barely visible above zero. MARY-Nano (blue, vertex only) is our subject-specific 224M-parameter model, included to illustrate the effect of per-subject adaptation (see Table 7).

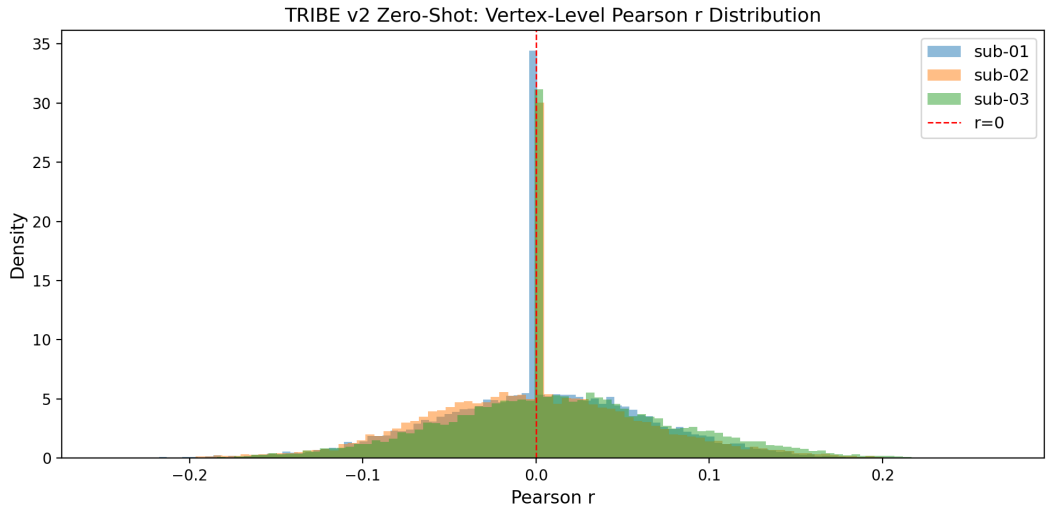

Figure 3. Distribution of vertex-level Pearson r for each subject. Sub-01 (blue) and sub-02 (orange) are tightly centered on zero with a sharp peak. Sub-03 (green) shows a slight rightward shift with a longer positive tail, consistent with its higher mean r in Table 4. All distributions span $r \in [-0.25, 0.25]$, indicating that even the best vertices achieve only modest correlations.

1. **Spatial misalignment**: The mapping from stimulus features to cortical locations is subject-specific. Averaging over misaligned spatial maps produces a blurred representation that is anti-correlated with any individual (demonstrated by the visual cortex result).
2. **Temporal dynamics mismatch**: The predictions have a lag-1 autocorrelation of 0.940, versus 0.12 for preprocessed BOLD (after detrending and z-scoring). The average embedding smooths over subject-specific hemodynamic response profiles, producing predictions that are temporally smooth but poorly synchronized with individual hemodynamic fluctuations.
3. **Response magnitude calibration**: Without per-subject adaptation, the model cannot calibrate the absolute magnitude of responses in each region. This is particularly problematic for regions with strong subject-specific gain differences (e.g., amygdala, ventral visual areas).

5.2 The Hierarchy of Transferability

The Yeo-7 results suggest a principled hierarchy of how well brain representations transfer across subjects without adaptation:

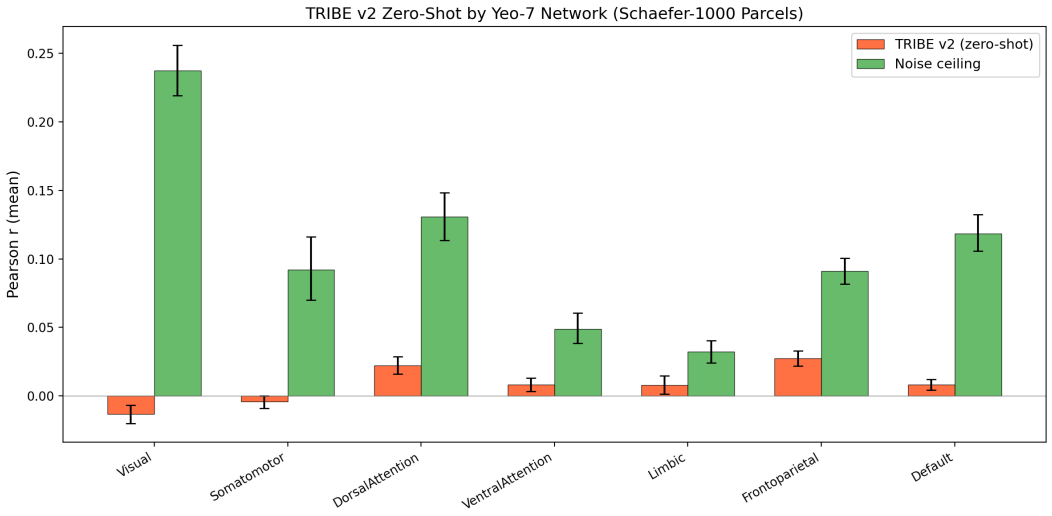

Figure 4. Yeo-7 network decomposition at Schaefer-1000 parcel level. Green bars: inter-subject noise ceiling. Orange bars: TRIBE v2 zero-shot. Visual cortex (highest ceiling) has *negative* TRIBE v2 prediction. Frontoparietal cortex (modest ceiling) has the best relative TRIBE v2 performance.

$$\text{Frontoparietal} > \text{Limbic} > \text{DorsalAttn} \approx \text{VentralAttn} > \text{Default} \gg \text{Visual, Somatomotor}$$

This ordering maps onto the cortical hierarchy from unimodal (sensory/motor) to transmodal (association) cortex [11]. Higher-order representations—semantic meaning, narrative comprehension, attentional allocation—are more consistent across individuals because they are abstracted away from idiosyncratic sensory processing. This has practical implications: zero-shot brain models may be viable for *cognitive* predictions (e.g., attention decoding, semantic categorization) even when they fail at *perceptual* predictions (e.g., retinotopic mapping, tonotopic organization).

5.3 Implications for Foundation Model Evaluation

These results carry a broader lesson for the neuroimaging community. "Foundation model" is borrowed from NLP, where models like GPT-4 and Llama demonstrate strong zero-shot performance across diverse tasks. In neuroimaging, the analogy is misleading. A brain-predictive foundation model is closer to a *pre-trained backbone* in computer vision (e.g., ImageNet-pretrained ResNet): useful as a feature extractor, but requiring task-specific fine-tuning to produce predictions.

We recommend that publications reporting brain-predictive foundation models include:

1. **Zero-shot baselines** alongside fine-tuned results, so readers can distinguish architectural contributions from adaptation contributions.
2. **Per-subject variability** in zero-shot performance, to quantify how well the average embedding centers the population.
3. **Measured noise ceilings** on the evaluation data, not literature estimates from different datasets or preprocessing pipelines.
4. **Network-level decomposition**, since global averages can mask systematic failure in specific functional systems.

5.4 Comparison with Subject-Specific Approaches

To contextualize these zero-shot results, we compare against TRIBE v2's own fine-tuned performance and include a subject-specific model from our lab as an additional reference point:

Model	Mean r	% Ceiling
TRIBE v2 zero-shot (vertex)	0.005	4.3%
MARY-Nano (vertex)†	0.055	46.7%
TRIBE v2 fine-tuned (parcel)	0.215	—

Table 7. Performance comparison. †MARY-Nano is our own 224M-parameter 6-stream model (SlowFast R101 backbone) trained directly on per-subject vertex-level data from the same CNeuroMod subjects. It is included to illustrate the effect of per-subject adaptation, not as a head-to-head architectural comparison. The parcel-to-vertex comparison with fine-tuned TRIBE v2 is not directly valid due to differing spatial resolutions.

The key observation is not the specific models involved but the magnitude of the gap between zero-shot and subject-adapted approaches. MARY-Nano, trained with per-subject data on the same subjects and stimulus, achieves 46.7% of the noise ceiling at vertex level—demonstrating that the signal *is* present in the data and recoverable with subject-specific adaptation. The fine-tuning step is not a "nice to have"; it is the mechanism that transforms a generic feature extractor into a functional brain model.

5.5 Practical Guidance for Practitioners

For researchers considering using TRIBE v2's released checkpoint:

1. **Do not use zero-shot predictions as brain maps.** The average embedding produces spatially incoherent predictions that are anti-correlated with individual subjects in sensory cortex.
2. **Use the checkpoint as a feature backbone.** Extract multimodal features from the intermediate layers and train per-subject linear probes on held-in fMRI data.
3. **Collect calibration data.** A minimum of 10–20 minutes of fMRI data per subject (watching a different stimulus) is likely necessary to learn the subject-specific projection.
4. **Evaluate per-network.** Global metrics mask systematic failures. Report Visual, Default, and Frontoparietal performance separately.

5.6 Limitations

1. **Single stimulus**: We evaluated on one 603-second movie segment. Performance may differ for other stimulus types (e.g., rest, task-based paradigms, different movie genres).
2. **Limited subjects**: Three subjects for vertex-level analysis (four for parcel-level). A larger sample would tighten the noise ceiling estimate and reduce inter-subject averaging noise.
3. **No HRF exploration**: We used linear interpolation without HRF convolution, following TRIBE v2's documentation. An exhaustive temporal alignment search (varying delays, HRF shapes) might recover marginally more signal.
4. **No fine-tuning comparison on same data**: We did not perform per-subject fine-tuning of TRIBE v2 on this data. Such a comparison would quantify the exact gain from adaptation but would require substantial additional fMRI scanning time to collect training data.
5. **Average embedding version**: The released checkpoint may be updated. Our results apply to the version available via `facebook/tribev2` as of May 2026.

6 Conclusion

TRIBE v2 is an impressive scientific contribution—its multimodal architecture and challenge-winning results represent a genuine advance in computational brain encoding. But the publicly released checkpoint, equipped with an average subject embedding, does not produce meaningful zero-shot brain predictions.

Across 3–4 subjects watching the Bourne Ultimatum, the average embedding achieves $r = 0.0051$ at vertex level (4.3% of noise ceiling). It is anti-correlated with visual cortex responses ($r = -0.013$)—the most reliably stimulus-driven region of the brain. It captures less than 30% of the noise ceiling in

any Yeo-7 network. The parcel-level scatter (Figure 1) shows zero relationship between a region's predictability and the model's prediction quality.

The "foundation model for the brain" is not a model of any brain. It is a pre-trained feature extractor. Its value lies in the quality of features it provides for downstream per-subject fine-tuning—not in its zero-shot predictions. Researchers and practitioners downloading the checkpoint should not expect usable predictions without investing in subject-specific adaptation data and training.

We release all measurement data, bootstrap confidence intervals, per-subject breakdowns, and publication-quality visualizations to serve as a reference zero-shot benchmark for the neuroimaging community.

A Computational Details

All experiments were conducted on a single NVIDIA H100 SXM 80GB GPU (RunPod on-demand, $3.29/hr). Total pod uptime was 95 minutes, with 65 minutes consumed by TRIBE v2 inference and 18 seconds by all analysis phases combined.

Phase	Time (s)
1. B2 data download	18.3
2. TRIBE v2 inference	3,913.1
3. Structure analysis	1.9
4. Pearson correlation	7.5
5. Noise ceiling	6.2
6. Yeo-7 breakdown	1.8
7. Publication plots	1.0
Total	**3,949.8**

Table 8. Per-phase timing. Analysis phases (3–7) together take 18.4 seconds—0.5% of total runtime. The cost is entirely dominated by TRIBE v2 inference.

Software versions: PyTorch 2.4.0, CUDA 12.4, tribev2 (git HEAD, May 2026), nilearn 0.11.x, scipy 1.14.x, matplotlib 3.9.x.

Bootstrap CIs used 10,000 resamples with a fixed seed (42) for reproducibility.

B Raw Prediction Statistics

Statistic	Value
Shape	$(1682, 20484)$
Data type	float32
Mean	0.024
Std	0.143
Min	-1.417
Max	1.489
Lag-1 autocorrelation	0.940

Table 9. Descriptive statistics of TRIBE v2 raw predictions before temporal resampling.

C Data and Code Availability

Raw TRIBE v2 predictions (137.8 MB, `.npy`), per-subject correlation maps, noise ceiling arrays, and all publication figures are archived on Backblaze B2 at prefix `orcle/benchmark/2026-05-20-v2/`. Analysis code is available at https://github.com/ygali04/orcle. The CNeuroMod fMRI data used in this study are publicly available under CC0 license [3].

Disclosure

MARY-Nano is developed by the author at BrainVI. Its inclusion in Table 7 and Figure 2 serves to illustrate the general effect of per-subject adaptation on prediction quality and is not intended as a formal head-to-head benchmark. The primary findings of this paper—the failure of zero-shot prediction and the network-level decomposition—are independent of any comparison model.

References

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

[2] R. Cichy, G. Roig, A. Oliva, et al. The Algonauts Project 2025: How the Human Brain Makes Sense of Multimodal Movies. *arXiv preprint arXiv:2508.10784*, 2025.

[3] J. Boyle, F. Paugam, A. Bhagwat, et al. The Courtois Project on Neuronal Modelling: 2020 Data Release. *Scientific Data*, 2024.

[4] Meta AI. V-JEPA2: Self-supervised Video Representation Learning with Joint-Embedding Predictive Architectures. *arXiv*, 2025.

[5] A. Baevski, Y. Zhou, A. Mohamed, M. Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *NeurIPS*, 2020.

[6] C. Eren, F. Javed, et al. VIBE: Video-Informed Brain Encoding for Naturalistic fMRI Prediction. *arXiv preprint arXiv:2509.01127*, 2025.

[7] A. Schaefer, R. Kong, E. Gordon, et al. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.

[8] B. Yeo, F. Krienen, J. Sepulcre, et al. The Organization of the Human Cerebral Cortex Estimated by Intrinsic Functional Connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.

[9] U. Hasson, Y. Nir, I. Levy, G. Fuhrmann, R. Malach. Intersubject Synchronization of Cortical Activity During Natural Vision. *Science*, 303(5664):1634–1640, 2004.

[10] E. Finn, X. Shen, D. Scheinost, et al. Functional Connectome Fingerprinting: Identifying Individuals Using Patterns of Brain Connectivity. *Nature Neuroscience*, 18(11):1664–1671, 2015.

[11] D. Margulies, S. Ghosh, A. Goulas, et al. Situating the Default-Mode Network Along a Principal Gradient of Macroscale Cortical Organization. *Proceedings of the National Academy of Sciences*, 113(44):12574–12579, 2016.