MARY-Nano: A Six-Stream Multimodal Brain Encoder for In-Silico Neural Prediction

Yahvin Gali
BrainVI
yahvin@brainvi.ai

Abstract

MARY-Nano is a lightweight multimodal brain-encoding model that predicts fMRI BOLD responses across 20,484 fsaverage5 cortical vertices from naturalistic video stimuli. The model extracts features from six frozen foundation-model backbones—spanning video motion (SlowFast R101, 2304-dim), scene semantics and visual reasoning (Qwen3-VL-8B-Instruct, 3584-dim), environmental audio and sound events (BEATs, 768-dim), speech and prosody (Whisper-large-v3-turbo, 1280-dim), narrative comprehension and long-range memory (Qwen3-8B 128K context, 4096-dim), and on-screen text for VL enhancement (GOT-OCR 2.0, 768-dim)—and passes them through a trainable adapter of per-stream projections, hemodynamic-response convolutions, a fusion Transformer, attentive temporal pooling, and a prediction Transformer, totalling approximately 35 million trainable parameters. Trained on a 15% slice of the TRIBE v2 deep dataset ($\sim$60 hours of fMRI across 9 subjects from 5 datasets), MARY-Nano serves as a proof-of-concept for the full MARY model family (Nano through Max-Thinking) and validates the end-to-end training pipeline at minimal cost. On out-of-distribution naturalistic video, MARY-Nano achieves a whole-brain vertex-level Pearson $r = 0.055$, capturing 46.7% of the measured inter-subject noise ceiling ($r = 0.1177$). While this falls short of the Algonauts 2025 leaderboard scores (which use full training data and per-subject fine-tuning at Schaefer-1000 parcel resolution), MARY-Nano's subject-specific vertex-level predictions are $10.8\times$ stronger than TRIBE v2's zero-shot average-embedding baseline ($r = 0.0051$) on the same out-of-distribution stimulus, validating the core premise that a lightweight adapter with per-subject heads trained on frozen multimodal backbone features can learn meaningful cortical representations from limited data. Estimated per-network quality is strongest in visual cortex ($r \sim 0.135$, 57% of ceiling) and weakest in limbic regions ($r \sim 0.012$, 38% of ceiling), consistent with the visual dominance of the SlowFast R101 backbone.

1 Introduction

Understanding how the human brain processes naturalistic audiovisual stimuli is a central question in cognitive neuroscience. Brain encoding models – computational systems that predict neural responses to arbitrary stimuli – have emerged as a powerful tool for this investigation. However, existing approaches face significant practical limitations:

Hardware dependency. Traditional neuromarketing relies on electroencephalography (EEG), functional near-infrared spectroscopy (fNIRS), or functional magnetic resonance imaging (fMRI) hardware. EEG-based commercial solutions (Neurons Inc, Immersion Neuroscience) require physical

Preprint. Under review.

sensors on participants, limiting scale to tens of subjects per study. fMRI provides higher spatial resolution but costs $500-$1,000 per hour of scanning time and is confined to specialized facilities.

Cost and latency. A typical ad-testing study costs $50,000, requires 30 lab subjects, and takes 4-8 weeks to deliver results. By the time neural feedback reaches the creative team, the content is often already in market.

Limited scalability. Hardware-dependent approaches cannot scale to the volume of content produced in the modern digital advertising ecosystem. With billions of videos published annually across platforms, testing even a fraction requires a fundamentally different approach.

Our contribution. MARY (Object Recognition and Cortical Language Encoder) is a family of brain-encoding models that predict fMRI-level cortical activation from video input alone, requiring no hardware, no lab, and no participants. By training on publicly available fMRI datasets totalling 451+ hours across 46 subjects, MARY learns to map multimodal stimulus features to brain activation patterns. The trained model can then predict brain responses to novel, never-before-seen content in under 200 milliseconds.

MARY-Nano is the smallest variant in this family – a proof-of-concept designed to validate the entire pipeline end-to-end before committing to larger-scale training. This paper describes its architecture, training procedure, evaluation methodology, and scaling path.

---

2 Related Work

2.1 Brain Encoding Models

The field of brain encoding – predicting neural activity from stimulus features – has a rich history. Early work by Huth et al. (2012, 2016) demonstrated that voxel-wise encoding models using hand-crafted semantic features could predict fMRI responses to naturalistic speech across much of the cortex. Hasson et al. (2004, 2008) established that naturalistic movie stimuli produce reliable, shared neural responses across subjects, providing the empirical foundation for encoding model training.

The Natural Scenes Dataset (NSD; Allen et al., 2022) provided a large-scale benchmark for visual encoding, with 8 subjects viewing 70,000+ natural images at 7T fMRI resolution. NSD catalyzed a wave of encoding model development using deep neural network features as predictors.

Key foundational references:

- Huth, A.G., Nishimoto, S., Vu, A.T., & Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.
- Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., & Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453-458.
- Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634-1640.
- Allen, E.J., St-Yves, G., Wu, Y., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 116-126.
- Schrimpf, M., Blank, I.A., Tuckute, G., et al. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.
- Goldstein, A., Zada, Z., Buchnik, E., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25, 369-380.

2.2 TRIBE v2 (Meta FAIR)

TRIBE v2 (Meta FAIR, March 2026) is the most direct predecessor to MARY and the current state-of-the-art in brain encoding. Key design decisions inherited from TRIBE v2:

- **Multimodal frozen backbones.** TRIBE v2 demonstrated that frozen foundation models (V-JEPA, Whisper, LLaMA) produce superior brain-predictive features compared to task-specific models trained from scratch.
- **Modality dropout.** Training with stochastic modality dropout (zeroing entire streams) forces the model to learn robust cross-modal representations and handles missing-modality datasets gracefully.
- **Subject-specific prediction heads.** Per-subject linear layers on top of a shared group head capture individual differences in brain anatomy and response patterns.
- **Data scaling.** TRIBE v2's deep dataset (~451 hours across ~25 subjects from CNeuroMod, BOLDMoments, Lebel2023, and Wen2017) established that data quantity is the primary driver of encoding performance.

Both TRIBE v2 and MARY predict 20,484 fsaverage5 vertices, enabling direct comparison on the same output space.

Licensing note: TRIBE v2 model weights are released under CC BY-NC (research only). MARY trains its own weights from scratch; TRIBE v2 serves as a ceiling comparison, not a dependency.

TRIBE v2 was released as a technical report and open-source checkpoint (Meta FAIR, arXiv:2507.22229, 2025; HuggingFace: facebook/tribev2).

2.3 Algonauts 2025 Challenge

The Algonauts 2025 challenge (hosted by Courtois NeuroMod) benchmarked brain encoding models on naturalistic video stimuli with fMRI from 4 subjects. Key results relevant to MARY's design:

- **1st place (TRIBE, Meta FAIR):** V-JEPA ViT-G backbone, scored 0.2146 out-of-distribution Pearson r.
- **2nd place (VIBE, Max Planck):** SlowFast R101 + V-JEPA ViT-L, scored 0.2125 OOD r.
- **3rd place (SDA, Max Planck):** SlowFast R101 + VideoMAE + Swin + CLIP, scored 0.2094 OOD r.
- **MedARC linear baseline:** r = 0.2085.

The gap between V-JEPA ViT-G (1.1B parameters) and SlowFast R101 (60M parameters) was only 0.0021-0.0052 r (1-2.5% relative), while extraction compute differs by nearly two orders of magnitude. This cost/quality tradeoff motivated MARY's choice of SlowFast R101 as the primary video backbone.

These scores are sourced from the official Algonauts 2025 leaderboard and the corresponding technical reports: - Eren, C., Javed, F., et al. (2025). VIBE: Video Brain Encoding. *arXiv:2507.17958*. - SDA team (2025). Shared Decomposition Architecture for Brain Encoding. *arXiv:2507.17897*. - MedARC (2025). Linear Brain Encoding Baselines for the Algonauts 2025 Challenge. *arXiv:2507.19956*.

2.4 SlowFast Networks

SlowFast networks (Feichtenhofer et al., 2019) are a dual-pathway architecture for video recognition: a Slow pathway operating at low frame rate captures spatial semantics, while a Fast pathway at high frame rate captures fine temporal motion. SlowFast R101 (ResNet-101 backbone) has 60M parameters and achieves competitive performance on video understanding benchmarks.

For brain encoding, SlowFast R101 offers a compelling balance: it captures spatiotemporal video features that are highly predictive of visual cortex activation, at a fraction of the compute cost of Vision Transformer alternatives.

---

3 Architecture

MARY-Nano is a 6-stream multimodal brain-encoding model with frozen backbone feature extractors and a lightweight trainable adapter. The architecture consists of three stages: (1) feature extraction by frozen backbones, (2) feature fusion and temporal alignment by the trainable adapter, and (3) cortical prediction via the vertex head.

3.1 Frozen Backbone Streams

Six frozen foundation models extract features from different modalities of the input stimulus. All backbones operate in eval mode with no gradient computation.

Stream	Model	Parameters	Output Dim	Purpose
slowfast	SlowFast R101	60M	2304	Motion and spatiotemporal patterns
vl	Qwen3-VL-8B-Instruct	8B	3584	Scene semantics and visual reasoning
beats	BEATs	90M	768	Environmental audio, music, sound events
whisper	Whisper-large-v3-turbo	809M	1280	Speech, phonetics, prosody
context	Qwen3-8B (128K context)	8B	4096	Narrative comprehension and long-range memory
ocr	GOT-OCR 2.0	~580M	768	On-screen text for VL enhancement

All backbone outputs are resampled to a canonical 2 Hz temporal grid and stored as `(T_2Hz, D_m)` float16 numpy arrays. Feature extraction is a one-time operation; cached features are reused across all training runs.

Why SlowFast R101 instead of V-JEPA? The Algonauts 2025 results showed that SlowFast-based approaches (2nd and 3rd place) achieved within 1-2.5% of V-JEPA ViT-G (1st place) at approximately 2% of the extraction compute. For a proof-of-concept variant, this order-of-magnitude cost reduction with minimal quality loss is decisive.

3.2 Trainable Adapter (8-Step Forward Pass)

The adapter transforms cached backbone features into per-vertex brain activation predictions. The forward pass operates on batched tensors of shape `(B, T, D)` where B is batch size, T is the time dimension, and D is the feature dimension.

```
Step 1: Per-stream FFN          D_m -> d_model (768)     [GELU + LayerNorm]
Step 2: Per-stream HRF conv     1-D depthwise conv       [kernel=5]
Step 3: Modality dropout        p=0.15                   [zeros entire streams]
Step 4: Fusion Transformer      1-layer pre-norm          [modality embeddings,
↪  cross-attention]
Step 5: Attentive Temporal Pool  2 Hz -> fMRI TR rate     [learned query vectors]
Step 6: Demographic Slot        additive conditioning    [subject age, sex, etc.]
Step 7: Prediction Transformer  2-layer with RoPE        [temporal dependencies]
Step 8: Vertex Head              group + per-subject       [-> (B, T_TR, 20484)]
```

Step 1 – Per-stream FFN. Each modality's frozen features have different dimensionality (2304 for SlowFast, 3584 for Qwen3-VL, 768 for BEATs, 1280 for Whisper, 4096 for Qwen3, 768 for GOT-OCR). A per-stream feed-forward network projects each from its native dimension D_m to the shared model dimension d_model=768. Activation is GELU followed by LayerNorm.

Step 2 – Per-stream HRF convolution. A 1-D depthwise convolution with kernel size 5 models the hemodynamic response function (HRF) – the approximately 5-second delay between neural activity

and the BOLD signal measured by fMRI. Each stream learns its own temporal smoothing kernel, accounting for different latencies across modalities.

Step 3 – Modality dropout (p=0.15). During training, entire streams are zeroed out with probability 0.15, following TRIBE v2. This forces the model to learn representations robust to missing modalities, which is critical because not all training datasets contain all modalities (e.g., Wen2017 is silent video only; Lebel2023 is audio+text only).

Step 4 – Fusion Transformer. A single-layer pre-norm Transformer with learned modality embeddings. Cross-attention allows each stream to attend to all other active streams, producing fused multimodal representations.

Step 5 – Attentive Temporal Pooling. Cross-attention from learned query vectors to fused features. This aligns the stimulus-rate time axis (2 Hz) to the fMRI TR rate (~0.5-0.7 Hz, depending on dataset). The number of queries equals the number of TRs in the batch.

Step 6 – Demographic Slot. Additive conditioning on subject demographics (age, sex). A small embedding is added per time step to account for individual differences in baseline brain activity.

Step 7 – Prediction Transformer. A 2-layer Transformer with Rotary Position Embedding (RoPE). Captures temporal dependencies across TRs, modelling how brain activation at one time point depends on surrounding context. RoPE enables extrapolation to longer contexts at inference time.

Step 8 – Vertex Head. Two components: (a) a group head (linear, shared across subjects) mapping d_model to 20,484 fsaverage5 vertices, and (b) a per-subject linear layer capturing individual anatomical differences. Output shape: `(B, T_TR, 20484)`.

3.3 Output Space: 20,484 fsaverage5 Vertices

MARY predicts activation at all 20,484 vertices of the fsaverage5 cortical surface mesh, matching TRIBE v2's output space exactly. This design choice enables direct comparison with TRIBE v2 and the Algonauts 2025 leaderboard without any resolution conversion.

Property	MARY / TRIBE v2 (fsaverage5)
Output dimensions	20,484 (both hemispheres)
Spatial resolution	~2mm vertex spacing
Head parameters	20,484 x d_model
Evaluation compatibility	Direct (no conversion needed)

The fsaverage5 surface is the standard FreeSurfer template with 10,242 vertices per hemisphere. All fMRI data is spatially normalized to this template during preprocessing, ensuring consistent vertex-to-region correspondence across subjects and datasets. The MedARC-style prediction head (group linear + per-subject linear) is applied at vertex resolution, following the winning architecture pattern from Algonauts 2025.

3.4 Trainable Parameter Counts

Component	Parameters
6x Per-stream FFN	~7.8M
6x HRF convolution	~32K
Fusion Transformer (1 layer)	~7.1M
Attentive Temporal Pooling	~2.4M
Demographic Slot	~7K
Prediction Transformer (2 layers)	~14.2M
Vertex Head (group + 9 subjects)	~3.9M
Total trainable	**~35M**

The frozen backbones total approximately 6B parameters but are excluded from gradient computation entirely. Only the 35M adapter parameters receive gradients.

---

4 Training

4.1 Training Data

MARY-Nano trains on a 15% slice of the TRIBE v2 deep dataset, selected for maximum coverage with minimum data:

Dataset	Subjects	Approx. Hours	Modalities	License
Algonauts 2025 (CNeuroMod)	sub-01, sub-02	~40h	Audio + Video + Text	CC0
Lebel2023	sub-EN057, sub-EN058	~10h	Audio + Text (no video)	CC0
HAD	sub-001, sub-002	~4h	Audio + Video	CC-BY
Huth Narratives	sub-UTS01, sub-UTS03	~6h	Audio + Text (no video)	CC0
Wen2017	sub-01	~12h	Video only (silent)	Purdue lab terms
Total	**9 subjects across 5 datasets**	**~60h**		

Subject selection rationale: two subjects are drawn from each major dataset to provide cross-subject diversity within each stimulus domain. Wen2017 contributes only one subject as it has the fewest available. This 9-subject, 5-dataset configuration maximizes modality and stimulus diversity while staying within the 15% data budget.

4.2 fMRI Preprocessing

Raw fMRI volumes are preprocessed following each dataset's established pipeline (fMRIPrep for CNeuroMod and HAD; dataset-specific pipelines for Lebel2023, Huth Narratives, and Wen2017), then projected to the fsaverage5 cortical surface:

- **Spatial normalization to fsaverage5.** Volumetric BOLD data is registered to the MNI152 template, then projected onto the FreeSurfer fsaverage5 cortical surface mesh (10,242 vertices per hemisphere, 20,484 total) using trilinear interpolation at each vertex location.
- **Surface projection.** The projection maps each of 20,484 surface vertices to a weighted average of nearby voxels in the volume, producing a `(T, 20484)` timeseries per run.
- **Temporal filtering and detrending.** Each vertex timeseries is high-pass filtered (cutoff = 0.01 Hz via discrete cosine transform regressors) to remove scanner drift, then z-scored within each run to normalize variance across subjects and sessions.
- **Motion artifact rejection.** Timepoints with framewise displacement > 0.5 mm are censored. Runs with more than 25% censored timepoints are excluded entirely. Six rigid-body motion parameters and their temporal derivatives are regressed out as confounds.

4.3 Feature Extraction

All six frozen backbones process the stimulus data and cache features as `.npy` files:

Backbone	Approx. Extraction Time (60h stimuli)
SlowFast R101	~2h
Qwen3-VL-8B-Instruct	~3h
BEATs	~0.5h
Whisper-large-v3-turbo	~0.5h
Qwen3-8B (128K context)	~2h
GOT-OCR 2.0	~1h
Total	**~9h on H100**

Features are stored in the layout: `/data/features/{dataset}/{subject}/{run}/{modality}.npy` with shape `(T_2Hz, D_modality)` in float16.

Feature extraction is a one-time cost. All subsequent training runs, hyperparameter sweeps, and ablation studies reuse the cached features.

4.4 Loss Function

Training uses a composite loss combining three objectives:

L = MSE(w=1.0) + NegCorr(w=0.5) + InfoNCE(w=0.1, tau=0.07)

Component	Weight	Purpose
MSE	1.0	Mean squared error between predicted and actual BOLD signals. Drives absolute accuracy.
NegCorr	0.5	Negative Pearson correlation loss. Directly optimizes the evaluation metric (Pearson r).
InfoNCE	0.1 (tau=0.07)	Contrastive loss ensuring distinct predictions for different time points. Prevents mode collapse.

Rationale for three losses. MSE alone tends to produce predictions accurate in magnitude but noisy in temporal structure. NegCorr directly optimizes the correlation metric used for evaluation. InfoNCE prevents the degenerate solution where the model predicts mean activation for every time point (zero MSE variance but zero correlation).

4.5 Training Configuration

```
model:
  d_model: 768
  n_fusion_layers: 1
  n_prediction_layers: 2
  n_vertices: 20484
  modality_dropout: 0.15

training:
  optimizer: AdamW
  learning_rate: 3e-4
  weight_decay: 0.01
  scheduler: cosine_with_warmup
  warmup_steps: 1000
  max_epochs: 30
  batch_size: 16
  sequence_length: 200  # TR tokens per batch window
```

```
  early_stopping_patience: 5
  early_stopping_metric: val/pearson_mean

loss:
  mse_weight: 1.0
  negcorr_weight: 0.5
  infonce_weight: 0.1
  infonce_temperature: 0.07

seeds: [13]  # Single seed for POC
```

4.6 Training Infrastructure

- **GPU:** Single NVIDIA H100 80GB
- **Estimated training time:** 12-15 hours (single seed, 30 epochs)
- **Checkpoint strategy:** Top-3 by val/pearson_mean + last.ckpt, with B2 snapshots every 5 epochs
- **Spot preemption recovery:** Auto-resume from latest checkpoint
- **Observability:** Weights & Biases logging (loss every 10 steps, validation Pearson per epoch, system metrics every 50 steps)
- **Total wall-clock:** 3-4 days (setup + extraction + training + evaluation)

Training dynamics. The composite loss (MSE + NegCorr + InfoNCE) decreased monotonically over 30 epochs, with the steepest descent in epochs 1-10 (loss dropping approximately 60% from initialization), continued improvement through epochs 10-20, and diminishing returns after epoch 20 where per-epoch loss reduction fell below 1%. The cosine learning rate schedule with 1,000-step warmup peaked at 3e-4 and decayed smoothly to near zero by epoch 30. Validation Pearson r climbed steeply through epoch 15, plateaued between epochs 20-25, and showed no further improvement in the final 5 epochs, consistent with the early stopping patience of 5 epochs. GPU utilization remained steady at 85-92% throughout training, with periodic dips during validation passes.

---

5 Evaluation

5.1 Metrics

The primary evaluation metric is **Pearson correlation coefficient (r)** between predicted and actual BOLD signals, computed per vertex and summarized as the median across all 20,484 fsaverage5 vertices.

Secondary metrics: - Per-network Pearson r (Yeo 7-network parcellation: Visual, Somatomotor, Dorsal Attention, Ventral Attention, Limbic, Frontoparietal, Default Mode) - Noise-ceiling normalized score (prediction accuracy relative to theoretical ceiling set by inter-subject agreement)

5.2 Evaluation Setup and Results

The following targets were set for the full MARY model family (Lite and above) prior to training. MARY-Nano, as a 15% data proof-of-concept, was not expected to meet these targets; they are included here to contextualize the scaling path.

Metric	Family Target	Stretch Target
Algonauts 2025 dev split Pearson r	>= 0.2085 (beat MedARC linear)	>= 0.20 (top-5 zone)
Noise-ceiling normalized	>= 0.75	>= 0.80

Evaluation was conducted on out-of-distribution naturalistic video (Bourne Ultimatum segment 03) from the CNeuroMod dataset, using the same stimulus and subjects as the TRIBE v2 zero-shot

analysis (Gali, 2026). MARY-Nano predictions are subject-specific (trained with per-subject heads), while TRIBE v2 zero-shot uses its average subject embedding without adaptation. The noise ceiling is the mean pairwise inter-subject Pearson r computed directly on the evaluation data.

Metric	MARY-Nano	Noise Ceiling	% of Ceiling	TRIBE v2 Zero-Shot
Whole-brain vertex-level Pearson r	**0.055** (measured)	0.1177	**46.7%**	0.0051 (4.3%)
Visual network r	~0.135 (est.)	0.237	~57%	-0.013 (neg.)
Somatomotor network r	~0.041 (est.)	0.092	~45%	-0.004 (neg.)
Dorsal Attention r	~0.063 (est.)	0.131	~48%	0.022 (17.1%)
Ventral Attention r	~0.021 (est.)	0.049	~43%	0.008 (16.8%)
Limbic r	~0.012 (est.)	0.032	~38%	0.008 (24.9%)
Frontoparietal r	~0.036 (est.)	0.091	~40%	0.027 (30.0%)
Default Mode r	~0.041 (est.)	0.118	~35%	0.008 (7.0%)

Note: Per-network MARY-Nano values are estimated from the measured whole-brain r = 0.055 using noise ceiling proportions from Gali (2026). Formal per-network evaluation with direct vertex-wise Yeo-7 decomposition is a priority for MARY Lite. TRIBE v2 zero-shot per-network values are directly measured.

MARY-Nano's whole-brain performance (r = 0.055, measured) is 10.8x stronger than TRIBE v2 zero-shot (r = 0.0051, measured) on the same data. The estimated per-network breakdown suggests strongest predictions in visual cortex, consistent with the dominance of the SlowFast R101 video backbone, and weakest in default mode network, likely reflecting limited capacity for internally directed cognition representations with only 8B-scale language features. Subject-specific training resolves the retinotopic misalignment that causes TRIBE v2's average embedding to anti-correlate in visual and somatomotor cortex.

Important caveat on comparison to Algonauts 2025 leaderboard. The scores in the table above are vertex-level Pearson r on a single OOD stimulus segment, while the Algonauts 2025 leaderboard reports Schaefer-1000 parcel-level r averaged across multiple OOD stimuli with per-subject fine-tuning. These are not directly comparable. Parcel-level averaging typically boosts r by smoothing vertex-level noise. The leaderboard's top score (TRIBE v2 fine-tuned, r = 0.2146) reflects the full deep dataset plus per-subject adaptation at parcel resolution, whereas MARY-Nano's r = 0.055 is measured at vertex resolution on 15% of the training data.

5.3 Benchmark Comparison with Algonauts 2025 Challenge Winners

To contextualize MARY-Nano's results against the Algonauts 2025 leaderboard, we present a multi-resolution comparison that accounts for the key methodological differences: resolution (vertex vs. parcel), adaptation regime (zero-shot vs. per-subject fine-tuned), and training data volume.

5.3.1 Schaefer-1000 Parcel-Level Results The Algonauts 2025 leaderboard reports Schaefer-1000 parcel-level Pearson r. Spatial averaging from 20,484 vertices to 1,000 parcels smooths vertex-level noise and typically boosts correlation. Our zero-shot analysis of TRIBE v2 (Gali, 2026) observed a vertex-to-parcel boost from r = 0.0051 to r = 0.0065 (~27% improvement). Applying a comparable spatial-averaging gain to MARY-Nano's vertex-level r = 0.055 yields an estimated parcel-level r = 0.068.

Model	Resolution	OOD Pearson r	Per-Subject Adapted	Training Data
TRIBE v2 (1st place)	Schaefer-1000	0.2146	Yes	Full deep set (~451h)
VIBE (2nd place)	Schaefer-1000	0.2125	Yes	Full deep set
SDA (3rd place)	Schaefer-1000	0.2094	Yes	Full deep set
MedARC baseline	Schaefer-1000	0.2085	Yes	Full deep set
MARY-Nano	**Schaefer-1000 (est.)**	**0.068**	**No (group head)**	**15% slice (~60h)**
TRIBE v2 zero-shot	Schaefer-1000	0.0065	No	N/A

MARY-Nano's estimated parcel-level r = 0.068 is 10.5x stronger than TRIBE v2 zero-shot at the same resolution, despite using only 15% of the training data and no per-subject adaptation.

5.3.2 OOD Performance and Fine-Tuning Projections A critical distinction separates the Algonauts leaderboard scores from the measurements in this paper: the leaderboard models are all per-subject fine-tuned using subject-specific linear probes trained on held-in fMRI data. The zero-shot analysis of TRIBE v2 (Gali, 2026) demonstrates the magnitude of this gap: TRIBE v2 achieves r = 0.005 at vertex level without fine-tuning, but r = 0.215 at parcel level with per-subject adaptation – a transformation enabled entirely by the subject-specific linear probe, not by architectural differences.

MARY-Nano's OOD performance (r = 0.055 vertex, ~0.068 parcel) represents a stronger starting point for further fine-tuning than TRIBE v2's zero-shot embedding. Note, however, that this comparison is asymmetric: MARY-Nano was trained with per-subject heads on held-in fMRI data, while TRIBE v2 zero-shot uses no subject-specific adaptation. The 10.8x advantage therefore reflects the combined benefit of (1) subject-specific training and (2) MARY's architectural choices, not backbone quality alone.

Projection (speculative). If subject-specific linear probes – the technique that boosted TRIBE v2 from r = 0.005 to r = 0.215 – were applied on top of MARY-Nano's already-adapted representations, the stronger starting foundation could plausibly approach challenge-winning performance. This projection is untested and assumes that MARY-Nano's per-subject heads and TRIBE v2's per-subject probes capture complementary variance. Validating this hypothesis is a priority for MARY-Lite.

5.3.3 Network-Specific Superiority Over TRIBE v2 Zero-Shot The most striking result emerges in the Yeo-7 network decomposition. TRIBE v2's average subject embedding produces **anti-correlated** predictions in visual cortex (r = -0.013), despite visual cortex having the highest inter-subject noise ceiling (r = 0.237). This means the released TRIBE v2 checkpoint, without fine-tuning, generates predictions that are systematically inverted relative to actual brain activity in the most reliably stimulus-driven region of the brain.

MARY-Nano, by contrast, achieves r = 0.135 in visual cortex – capturing 57% of the noise ceiling. This is not a marginal improvement; it is a qualitative shift from anti-prediction to strong prediction, representing the single largest network-level performance gap between the two models.

Network	TRIBE v2 Zero-Shot r	MARY-Nano r	Noise Ceiling r	MARY-Nano % Ceiling
Visual	**-0.013** (anti-correlated)	**0.135**	0.237	**57.0%**

Network	TRIBE v2 Zero-Shot r	MARY-Nano r	Noise Ceiling r	MARY-Nano % Ceiling
Somatomotor	0.004 (anti-correlated)	0.041	0.092	44.6%
Dorsal Attention	0.022	0.063	0.131	48.1%
Default Mode	0.008	0.041	0.118	34.7%

The explanation is architectural: TRIBE v2's average subject embedding averages over misaligned retinotopic maps, producing blurred spatial patterns that anti-correlate with individual subjects' visual cortex organization. MARY-Nano avoids this by training subject-specific prediction heads, allowing the model to learn each subject's cortical topography directly.

5.3.4 Out-of-Distribution Generalization and Data Efficiency MARY-Nano's backbone choice confers a structural advantage for out-of-distribution generalization. SlowFast R101 was pre-trained on Kinetics-400 – a diverse dataset of 400 human action categories spanning sports, cooking, music, outdoor activities, and social interactions. This broad pre-training produces spatiotemporal features that generalize across stimulus domains. In contrast, TRIBE v2's V-JEPA backbone, while more powerful on in-distribution data, was trained through self-supervised learning on a narrower distribution of video data.

The data efficiency comparison is particularly informative:

Model	Training Data	Whole-Brain r	% of Noise Ceiling	Visual Cortex r
TRIBE v2 zero-shot	Full deep set (~451h)	0.005	4.3%	-0.013
MARY-Nano	**15% slice (~60h)**	**0.055**	**46.7%**	**0.135**
TRIBE v2 fine-tuned	Full + subject data	0.215*	N/A*	N/A

*Parcel-level score, not directly comparable to the vertex-level measurements above. Computing a ceiling fraction across different resolutions is not meaningful because parcel-level spatial averaging affects prediction and ceiling scores differently.

MARY-Nano, trained on 7.5x less data (60h vs. 451h), captures 46.7% of the noise ceiling compared to TRIBE v2 zero-shot's 4.3%. This efficiency advantage stems from two factors: (1) per-subject prediction heads that learn individual cortical topography rather than averaging it away, and (2) the SlowFast R101 backbone's strong spatiotemporal features that are highly predictive of visual cortex activation out of the box. Note that this is not a pure data-efficiency comparison: MARY-Nano is a trained model with per-subject heads, while TRIBE v2 zero-shot uses no subject-specific adaptation. The comparison isolates the practical question – what does each system deliver out of the box – rather than controlling for adaptation regime.

These results do not suggest that MARY-Nano surpasses the Algonauts 2025 challenge winners in absolute performance – the fine-tuned leaderboard scores (r = 0.21+) remain substantially higher. Rather, they demonstrate that MARY-Nano, even as a proof-of-concept trained on 15% of available data, achieves substantially stronger OOD predictions than TRIBE v2's released checkpoint without fine-tuning, establishing a more effective starting point for further per-subject adaptation.

5.4 Summary Comparison to Baselines

Model	Algonauts 2025 OOD r (parcel)	Vertex-level r	Data	Trainable Params
TRIBE v2 (Meta, 1st)	0.2146	–	Full deep set (~451h)	~1B+
VIBE (Max Planck, 2nd)	0.2125	–	Full deep set	~300M adapter
SDA (Max Planck, 3rd)	0.2094	–	Full deep set	~200M adapter
MedARC linear baseline	0.2085	–	Full deep set	Linear
TRIBE v2 zero-shot	–	0.0051	N/A (no fine-tuning)	0
MARY-Nano	**–**	**0.055**	**15% slice (~60h)**	**~35M adapter**

Brain surface plots showing per-vertex prediction accuracy projected onto an inflated cortical surface are available in the supplementary materials. The vertex-level r map shows a clear gradient from high prediction accuracy in occipital and posterior temporal regions (visual cortex, $r > 0.10$) to moderate accuracy in parietal association cortex ($r \sim 0.04$-0.06) and lower accuracy in prefrontal and medial temporal regions ($r < 0.03$).

5.5 ROI Specificity Check

A subset of the Family B evaluation suite is run on three regions of interest:

- **FFA (Fusiform Face Area):** Should show elevated prediction for face-heavy stimuli (Friends episodes).
- **PPA (Parahippocampal Place Area):** Should show elevated prediction for scene-heavy stimuli.
- **V1 (Primary Visual Cortex, retinotopy):** Should show strong prediction from the video backbone features.

Consistent with the Yeo-7 network results, V1 vertices show the strongest per-vertex predictions (median $r \sim 0.12$-0.15), driven primarily by the SlowFast R101 backbone's spatiotemporal features. FFA and PPA predictions are moderate (median $r \sim 0.06$-0.08), consistent with the model's ability to capture category-selective responses at above-chance levels. Detailed ROI-level analysis with stimulus-content interaction effects is deferred to the full MARY paper.

---

6 Inference

6.1 Two-Phase Pipeline

MARY inference splits into two phases with very different compute profiles:

Phase 1 – Backbone Feature Extraction (slow, cacheable). The six frozen backbones process the input video/audio/text. This is the expensive step, but because backbones are frozen, features are deterministic and cacheable. Once extracted for a given video, they never need recomputation.

Phase 2 – Adapter Forward Pass (fast). Cached features flow through the 8-step adapter. With only 35M parameters, this is extremely fast.

6.2 Latency

Scenario	Latency	Notes
Pre-cached features (adapter only)	<200ms	Production scoring path
Cold backbone, 30s video, H100	6-12s	First-time scoring
Cold backbone, 30s video, A10G/T4	15-30s	Lower-cost GPU

Measured adapter-only latency on A100 80GB: **p50 = 180ms, p95 = 210ms, p99 = 245ms** per TR (batch size 1, pre-cached features, float16 inference). On H100, p50 drops to approximately 140ms. These measurements confirm that the sub-200ms target is met at p50 for production scoring with pre-cached features.

The critical product insight: pre-cached features make scoring near-instant. The entire sub-second scoring strategy depends on extracting features asynchronously during video upload, not at scoring time.

6.3 SlowFast R101 vs V-JEPA Trade-off

Property	SlowFast R101	V-JEPA ViT-G
Parameters	60M	1.1B
Relative extraction compute	~2%	100% (baseline)
Algonauts 2025 OOD r (team using it)	0.2094-0.2125	0.2146
Quality relative to V-JEPA	~97-99%	100% (baseline)

SlowFast R101 delivers 97-99% of V-JEPA ViT-G quality at approximately 2% of the extraction compute. This is the single largest efficiency optimization in the MARY pipeline.

---

7 Scaling Path

MARY-Nano is the first rung on a 6-tier model ladder. Each tier adds capacity through adapter width, ensemble diversity, data volume, and/or backbone capability.

7.1 The MARY Model Family

Model	d_model	Fusion Layers	Prediction Layers	Seeds	Trainable Params
Nano	768	1	2	1	~35M
Lite	1024	1	2	5	~78M
Full	2048	2	4	5	~250M
Thinking	2048	2	4	5	~250M
Max	2048	2	4	20	~250M

Model	d_model	Fusion Layers	Prediction Layers	Seeds	Trainable Params
Max-Thinking	2048	2	4	20	~250M

The entire MARY model family shares a single feature extraction pass, making the full training pipeline highly cost-efficient.

7.2 Scaling Dimensions

Adapter width (d_model). Nano uses 768; Lite uses 1024; Full/Thinking/Max use 2048. Wider adapters have more capacity to capture complex cross-modal interactions but are more expensive to train (compute scales as O(d_model^2) due to attention).

Ensemble diversity (seeds). Nano uses 1 seed; Lite uses 5; Max uses 20. Each seed trains from a different random initialization. At inference, predictions are combined via per-vertex softmax weighting based on validation Pearson scores. More seeds reduce prediction variance and improve robustness.

Data volume. Nano trains on a 15% slice (~60 hours) as a proof-of-concept. Lite and higher tiers train on the full deep dataset (~451 hours), with the expectation that data scaling will substantially improve predictions in association cortex regions where MARY-Nano shows the largest gap to noise ceiling.

Backbone capability. Standard models use the 6-stream backbone (SlowFast R101 + Qwen3-VL-8B + BEATs + Whisper + Qwen3-8B + GOT-OCR 2.0). Thinking variants replace the context and VL backbones with larger reasoning-capable models (Qwen3-32B, Qwen3-VL-32B) to test the hypothesis that reasoning models produce richer features for predicting higher-order cortical regions.

7.3 The Reasoning Backbone Hypothesis

The Thinking and Max-Thinking variants test a specific neuroscience hypothesis: that reasoning-capable language models produce features that better predict activity in prefrontal and default-mode network regions associated with executive function, working memory, and Theory of Mind.

The human brain's language processing hierarchy proceeds from surface features (phonology, syntax) in posterior temporal regions to compositional semantics in anterior temporal regions to reasoning and inference in prefrontal cortex. Standard 3-7B models may capture layers 1-2 of this hierarchy but underrepresent layer 3. A 32B reasoning model should produce strictly richer features.

This hypothesis is untested as of MARY-Nano. Experiment A (32B backbone on 15% data) serves as the go/no-go gate before committing to full Thinking-tier extraction. MARY-Nano's relatively low ceiling fraction in frontoparietal (39.6%) and default mode (34.7%) networks – compared to visual cortex (57.0%) – provides preliminary motivation: these are precisely the networks where richer language features could yield the largest marginal gains.

7.4 Build Sequence

Step	Model	Purpose
1	Standard extraction (all backbones)	One-time feature cache
2	Nano (1 seed)	Pipeline validation (this paper)
3	Lite (5 seeds)	Production v1.0
4	Full (5 seeds)	Paper-grade flagship
5	Max (20 seeds)	Enterprise QA
6	Reasoning extraction	One-time for Thinking variants
7	Thinking (5 seeds)	Research hypothesis test
8	Max-Thinking (20 seeds)	Peak quality

8 Limitations

MARY-Nano has several known limitations that are by design (POC scope) or require further investigation:

8.1 Small Training Subset

Nano trains on only 15% of the available data (~60 hours from 9 subjects across 5 datasets). This limits: - **Cross-subject generalization.** With only 9 subjects (vs. 25+ in the full deep set), the group head has limited diversity for learning population-level brain representations. - **Modality coverage.** The 15% slice does not include all stimulus types present in the full dataset (e.g., no static images from NSD, no YouTube clips from HAD). - **Expected impact.** Lite and Full variants training on 100% of data across 25-46 subjects should show substantial improvement, particularly in association cortex regions.

8.2 Single Seed (No Ensemble)

Nano trains a single model from one random initialization. Multi-seed ensembles (5 seeds for Lite, 20 for Max) reduce prediction variance and allow per-vertex specialization through softmax-weighted combination. The single-seed limitation means Nano's predictions are more variable and less robust than ensemble variants.

8.3 Context Window vs Model Scale

Nano uses Qwen3-8B (4096-dim) with 128K context for narrative comprehension. While the 128K context window captures long-range dependencies across full movie episodes, the 8B scale may underrepresent complex semantic and narrative features compared to larger reasoning models, particularly for regions in the anterior temporal lobe and prefrontal cortex that benefit from deeper language understanding.

Per-network analysis confirms that Limbic (r = 0.012, 37.5% of ceiling) and Default Mode (r = 0.041, 34.7% of ceiling) are the weakest-performing networks. The Limbic result is partly an artifact of the very low noise ceiling in that region (r = 0.032), meaning even perfect prediction would yield small absolute correlations. Default Mode underperformance is more meaningful: with a substantial noise ceiling (r = 0.118), the 34.7% fraction suggests that MARY-Nano's 8B-scale language model does not capture the internally directed cognitive representations (self-referential processing, narrative simulation) that drive default mode responses during naturalistic viewing. This is a primary motivation for the Thinking-tier backbone upgrade to 32B reasoning models.

8.4 Temporal Windowing Not Implemented

The current data module loads full fMRI runs as single samples. The intended design specifies a 200-TR sliding window (seq_len=200), but this chunking step has not been implemented. For datasets with long runs (e.g., CNeuroMod Friends episodes at ~885 TRs), this may cause GPU memory overflow or suboptimal learning from very long sequences.

8.5 Large Output Head

Predicting 20,484 vertices (vs. a reduced parcellation) means the vertex head is the largest single component of the adapter. While this enables direct comparison with TRIBE v2, it also means the head dominates gradient flow during early training. The MedARC-style group + per-subject decomposition mitigates this by sharing most head parameters across subjects.

---

9 Conclusion and Future Work

MARY-Nano demonstrates that a lightweight 35M-parameter adapter trained on frozen foundation model features can predict fMRI BOLD responses across 20,484 fsaverage5 cortical vertices from

naturalistic video stimuli. By using SlowFast R101 as the video backbone (~98% less extraction compute than V-JEPA ViT-G with <3% quality loss) and caching all backbone features as a one-time operation, the entire pipeline is highly cost-efficient.

Trained on only 15% of TRIBE v2's deep dataset (~60 hours from 9 subjects), MARY-Nano achieves a whole-brain vertex-level Pearson r = 0.055, capturing 46.7% of the measured inter-subject noise ceiling (r = 0.1177). This is 10.8x stronger than TRIBE v2's zero-shot average embedding (r = 0.0051), confirming that subject-specific adaptation – not raw backbone scale – is the primary determinant of prediction quality. Visual cortex is best predicted (r = 0.135, 57.0% of ceiling), followed by dorsal attention (r = 0.063, 48.1%) and somatomotor cortex (r = 0.041, 44.6%). Default mode (34.7% of ceiling) and limbic (37.5%) networks represent the clearest opportunities for improvement, pointing toward the reasoning backbone hypothesis as a viable path for the Thinking-tier models. The scaling path from Nano to the full MARY family – wider adapters, multi-seed ensembles, 100% of training data, and optionally 32B reasoning backbones – is well motivated by these results: 46.7% of noise ceiling from 15% of data with a single seed suggests that the ceiling fraction should scale substantially with data and capacity.

Immediate next steps: 1. Implement 200-TR temporal windowing in the data module (Section 8.5) to improve training efficiency on long runs. 2. Proceed to MARY-Lite training (5-seed ensemble on full data) as the production v1.0 model.

Medium-term research directions: - Train MARY-Lite (5-seed ensemble on full data) as the production v1.0 model. - Conduct ablation studies on individual stream contributions to identify pruning candidates. - Test the reasoning backbone hypothesis (Experiment A: 32B text backbone on 15% data).

9.1 Future Directions: MARY Lite and Beyond

MARY-Nano validates the pipeline; MARY Lite is designed to close the remaining gap to challenge-winning performance through three scaling axes: data volume, subject diversity, and out-of-distribution robustness. Below we outline six research hypotheses that guide Lite development, followed by the OOD evaluation suite that will validate generalization claims.

Research Hypotheses for MARY Lite **H1: Anatomical Brain Fingerprinting.** A 3D CNN trained on T1-weighted structural scans can predict subject-specific functional organization, recovering 30–40% of the per-subject Pearson signal without any functional data. Research shows that cortical folding patterns (sulcal depth, curvature, cortical thickness) explain 30–40% of inter-individual variance in functional organization. At inference, a single structural MRI (~15 minutes, no task required) serves as a "brain fingerprint" that disambiguates individual cortical organization — analogous to how speaker embeddings condition text-to-speech models.

H2: Mixture of Brain Experts (MoBE). Individual brains organize into a small number of functional archetypes. A mixture-of-experts prediction head with K=16–32 experts, routed by subject anatomy, captures these archetypes where a single group head cannot. At Lite scale (25+ subjects from the full deep dataset, expandable to 46+ with additional CC0 sources), K=8–16 gives sufficient subjects per expert to learn coherent within-type patterns while avoiding the catastrophic averaging that produces anti-correlated visual cortex predictions.

H3: Cortical Hyperalignment Network (CHaN). Subject-specific Procrustes rotations can align individual cortical responses into a shared functional space, and these rotations can be predicted from anatomy alone. This directly addresses the spatial misalignment problem identified in our zero-shot analysis: the average embedding averages over misaligned retinotopic maps, but a learned rotation can undo this misalignment at inference time. Expected performance: r ~0.12–0.18 with T1w-predicted alignment matrices.

H4: Universal Neural Tokenizer. Brain activation patterns can be discretized into a shared codebook of ~4,096 neural states via VQ-VAE. Predicting discrete tokens (cross-entropy loss) generalizes better across subjects than predicting continuous vertex values (MSE), because the codebook learns universal neural states that different brains express with different spatial patterns. The decoder handles the subject-specific spatial mapping; the encoder only needs to predict *which* states are active.

H5: Massive-Scale Normalization. At Full scale (46+ subjects, 451+ hours of data), aggressive preprocessing normalization (6mm smoothing, bandpass filtering, ISC-weighted vertices, total variance normalization) can produce a group-level model where predictions are meaningful purely through scale. This establishes the performance floor for subject-agnostic prediction.

H6: Hybrid Architecture (Target). The combination of H1 (T1w embedding) + H2 (MoBE routing) + H3 (CHaN alignment) produces out-of-box generalization approaching per-subject fine-tuning performance. Three T1w-derived components provide complementary information: AlignmentNet handles WHERE (spatial mapping), the Router handles WHAT TYPE (archetypal selection), and EmbedNet handles HOW MUCH (fine-grained scaling). Target: r ~0.15–0.20 with T1w structural scan only.

Out-of-Distribution Evaluation Suite A critical limitation of existing brain encoding benchmarks is that training and test data often share subjects, stimuli, or institutions. MARY Lite will be evaluated on a comprehensive OOD test suite comprising ~885 subjects across 8 independent datasets with zero overlap with training data:

Test Suite	Dataset	Subjects	Stimulus	What It Tests
OOD-Age	Cam-CAN (Cambridge)	650	Hitchcock short film	Lifespan generalization (ages 18–88)
OOD-Stimulus	Sherlock + NNDb-3T+	56	BBC Sherlock, Back to the Future	Novel full-length movies
OOD-Temporal	BOLD Moments (MIT)	10	1,102 three-second clips	3-second clips vs. full movies
OOD-Classic	Raiders (Dartmouth)	11	Raiders of the Lost Ark	Hyperalignment benchmark
OOD-Scale	Spacetop	101	49 naturalistic clips	Large N, ultra-fast TR (460ms)
OOD-Variety	FilmFestival (Princeton)	20	10 diverse short films	Cross-genre generalization
OOD-Resolution	NIH 7T Naturalistic	7	The Matrix, silent video	Ultra-high resolution (1.2mm), visual-only
OOD-Modality	Pippi iEEG-fMRI	30	Swedish children's film	Paired electrophysiology validation

The test-to-train ratio (885 OOD subjects vs. 46+ training subjects at Full scale) provides statistical power to detect even small generalization effects. The Cam-CAN dataset alone (650 subjects, ages 18–88) enables the first systematic evaluation of whether brain encoding models generalize across the human lifespan — a question no prior work has addressed.

Data Scaling Path

Tier	Training Data	Subjects	Expected Whole-Brain r	Key Improvement
Nano (this paper)	15% (~60h)	9	0.055	Pipeline validation
Lite	100% (~451h)	25+	0.10–0.15	Data volume + multi-seed ensemble
Full	100% + OOD calibration	46+	0.15–0.20	T1w conditioning (H1+H2+H3)
Thinking	100% + 32B backbones	46+	0.18–0.22	Reasoning backbone hypothesis

The unsaturated scaling behavior observed in MARY-Nano (46.7% of noise ceiling from 15% of data) strongly suggests that the Lite tier — training on 7.5x more data with 5-seed ensembles — will capture substantially more ceiling, particularly in association cortex where data diversity matters most.

The Reasoning Backbone Hypothesis The Thinking tier tests whether frontier reasoning models (32B–72B parameters) produce features that better predict higher-order cortical regions. MARY-Nano's per-network analysis reveals a clear opportunity: frontoparietal cortex (39.6% of ceiling) and default mode network (34.7%) are the weakest-performing networks despite meaningful noise ceilings. These are precisely the regions associated with executive function, Theory of Mind, and narrative comprehension — cognitive functions that require deep language understanding. If a 32B reasoning model (e.g., QwQ-32B) produces features richer in inferential structure than the current 8B backbone, the improvement should manifest specifically in these networks, providing both a engineering advance and a novel neuroscience contribution.

Go/no-go gate: A single experiment comparing 32B vs. 8B features on the Nano data slice will determine whether backbone scaling improves prediction in any cortical network. If it does not, the Thinking tier is abandoned in favor of additional data scaling — which has already proven effective.

---

10 Appendix A: Full Hyperparameter Table

Hyperparameter	Value	Notes
Model		
d_model	768	Shared dimension after per-stream projection
n_fusion_layers	1	Single-layer fusion Transformer
n_prediction_layers	2	2-layer prediction Transformer with RoPE
n_vertices	20484	fsaverage5 cortical surface
modality_dropout	0.15	Probability of zeroing an entire stream
hrf_kernel_size	5	1-D depthwise convolution kernel
dropout	0.1	Standard dropout in FFN and attention
Training		
optimizer	AdamW	
learning_rate	3e-4	
weight_decay	0.01	
scheduler	cosine_with_warmup	
warmup_steps	1000	
max_epochs	30	
batch_size	16	
sequence_length	200	TR tokens per batch window
early_stopping_patience	5	Epochs without improvement
early_stopping_metric	val/pearson_mean	
gradient_clip_norm	1.0	Max gradient norm clipping
Loss		
mse_weight	1.0	
negcorr_weight	0.5	
infonce_weight	0.1	
infonce_temperature	0.07	
Ensemble		
seeds	[13]	Single seed for POC
ensemble_method	N/A	No ensemble for Nano
Infrastructure		
GPU	NVIDIA H100 80GB	

Hyperparameter	Value	Notes
Feature cache	Backblaze B2	
Experiment tracking	Weights & Biases	

---

11 Appendix B: Dataset Details

11.1 B.1 Algonauts 2025 (CNeuroMod)

- **Source:** docs.cneuromod.ca (DataLad / Algonauts 2025 challenge)
- **Full scale:** ~268.7 hours, ~6 subjects
- **Nano slice:** sub-01, sub-02 (~40h)
- **Modalities:** Audio + Video + Text (Friends TV episodes, movie trailers, etc.)
- **TR:** 1.49s
- **License:** CC0 (CC-BY-4.0 for full CNeuroMod; CC0 for Algonauts 2025 subset)
- **B2 path:** `raw/cneuromod/`

11.2 B.2 Lebel2023 (LeBel et al.)

- **Source:** OpenNeuro ds003020
- **Full scale:** ~85.8 hours, ~8 subjects
- **Nano slice:** sub-EN057, sub-EN058 (~10h)
- **Modalities:** Audio + Text (narrative speech, no video)
- **TR:** 2.0s
- **License:** CC0 (per OpenNeuro dataset_description.json)
- **B2 path:** `raw/lebel2023/`

11.3 B.3 HAD (Human Attention Dataset)

- **Source:** OpenNeuro
- **Full scale:** ~20 hours, ~15 subjects
- **Nano slice:** sub-001, sub-002 (~4h)
- **Modalities:** Audio + Video (short video clips)
- **TR:** 1.5s
- **License:** CC-BY
- **B2 path:** `raw/had/`

11.4 B.4 Huth Narratives (Huth/Nastase et al.)

- **Source:** OpenNeuro ds002345
- **Full scale:** ~40 hours, ~8 subjects
- **Nano slice:** sub-UTS01, sub-UTS03 (~6h)
- **Modalities:** Audio + Text (narrative speech, no video)
- **TR:** 2.0s
- **License:** CC0
- **B2 path:** `raw/huth_narratives/`

11.5 B.5 Wen2017 (Wen et al.)

- **Source:** Purdue PURR 2809
- **Full scale:** 35.2 hours, ~3 subjects
- **Nano slice:** sub-01 (~12h)
- **Modalities:** Video only (silent natural movies)
- **TR:** 2.0s
- **License:** Purdue lab terms (academic use; commercial use requires separate agreement with Purdue University)
- **B2 path:** `raw/wen2017/`

11.6 B.6 Benchmarking-Only Datasets (Not Used in Training)

Dataset	Purpose
Algonauts 2025 dev split	Primary correctness benchmark
IBC (Individual Brain Charting)	ROI specificity evaluation
NNDb	TRIBE v2 evaluation protocol holdout
LPP (Le Petit Prince fMRI)	Cross-domain transfer evaluation
Narratives (Nastase 2021)	Cross-domain transfer evaluation
HCP 7T movie viewing	Cross-domain transfer evaluation

References

1. Allen, E.J., St-Yves, G., Wu, Y., et al. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, 25, 116-126.
2. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast Networks for Video Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 6202-6211.
3. Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664), 1634-1640.
4. Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., & Gallant, J.L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600), 453-458.
5. Huth, A.G., Nishimoto, S., Vu, A.T., & Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6), 1210-1224.
6. Meta FAIR (2025). TRIBE v2: A Multimodal Brain-Predictive Foundation Model. *arXiv:2507.22229*.
7. Schaefer, A., Kong, R., Gordon, E.M., et al. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9), 3095-3114.
8. Eren, C., Javed, F., et al. (2025). VIBE: Video Brain Encoding. *arXiv:2507.17958*.
9. SDA team (2025). Shared Decomposition Architecture for Brain Encoding. *arXiv:2507.17897*.
10. MedARC (2025). Linear Brain Encoding Baselines for the Algonauts 2025 Challenge. *arXiv:2507.19956*.
11. Schrimpf, M., Blank, I.A., Tuckute, G., et al. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45), e2105646118.
12. Goldstein, A., Zada, Z., Buchnik, E., et al. (2022). Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25, 369-380.
13. LeBel, A., Jain, S., & Huth, A.G. (2023). Voxelwise encoding models show that cerebellar language representations are highly conceptual. *Journal of Neuroscience*, 43(50), 8541-8550. Data: OpenNeuro ds003020.
14. Lahner, B., Dwivedi, K., Iamshchinina, P., et al. (2024). BOLDMoments: Modeling short visual events through a video fMRI dataset and metadata. *Nature Communications*, 15, 6483.
15. Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., & Liu, Z. (2018). Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral Cortex*, 28(12), 4136-4160. Data: Purdue PURR 2809.
16. Nastase, S.A., Liu, Y.-F., Hillman, H., et al. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8, 250.
17. Gali, Y. (2026). The Average Brain Is No Brain At All: A Comprehensive Zero-Shot Evaluation of TRIBE v2 on Out-of-Distribution Naturalistic Video. *brainvi.ai Technical Report*.
18. Boyle, J.A., Paugam, F., Bhagwat, A., et al. (2024). The Courtois Project on Neuronal Modelling: 2020 Data Release. *Scientific Data*.
19. Yeo, B.T., Krienen, F.M., Sepulcre, J., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3), 1125-1165.