

MultiFlow: Vision-Driven Multimodal fMRI Encoding via Stimulus-Informed Source Flow Matching

Bui Van Hung¹
hungbv16@jnu.ac.kr

Hqang-Son Vo¹
hoangsonvothanh@jnu.ac.kr

Mai Duc Minh Huy²
minhhuymaiduc@gmail.com

Dang Thanh Vu³
dtvu@aiseed.kr

Soo-Hyung Kim¹
shkim@jnu.ac.kr

¹ Chonnam National University
 Gwangju, South Korea

² Ho Chi Minh University of Science
 Ho Chi Minh City, Vietnam

³ AISEED, Inc.
 South Korea

Preprint. Accepted at the British Machine Vision Conference (BMVC) 2026.

Abstract

Current fMRI encoding models frame prediction as deterministic regression, yet identical stimuli evoke a distribution of neural responses (arising from physiological noise, hemodynamic variability, and cognitive fluctuations) that a single point estimate cannot capture. This work introduces **MultiFlow**, a generative fMRI encoding framework based on **Conditional Flow Matching (CFM)** and a formulation termed **Stimulus-Informed Source Flow Matching (SISFM)**. SISFM generalises CFM by conditioning the source distribution on the stimulus: a learnable temporal convolution shapes the source mean with an HRF-length receptive field, while a variance head shrinks its scale, shortening the expected transport path. SISFM is instantiated with a Diffusion Transformer (DiT-X) decoder parameterising the velocity field, with sub-layer Attention Residuals, heterogeneous fusion of ten frozen vision, audio, language, and audio-visual backbones, and per-subject Yeo-network heads. On the Algonauts 2025 dataset, a single MultiFlow checkpoint recovers most of the benefit of large deterministic ensembles: it trails the 1000-model TRIBE ensemble in-distribution by 0.003 PCC (0.3163 vs. 0.3195), and the strongest in- and out-of-distribution ensembles (VIBE-20, MedARC-50) by 0.010 and 0.009 PCC, using a single trained model rather than an ensemble of 20–1000. The same checkpoint also provides stimulus-conditioned uncertainty estimates and sample-based diversity.

1 Introduction

Naturalistic paradigms in which participants watch movies during fMRI scanning have become a standard setting for studying how the brain integrates vision, language, and sound

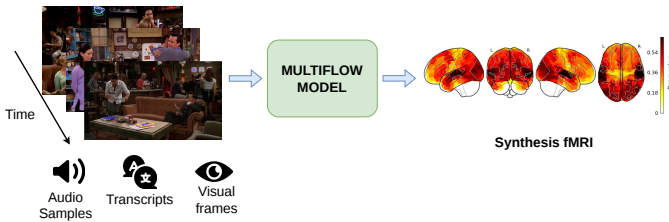


Figure 1: **Overview of the MultiFlow pipeline.** Time-aligned *visual frames*, *audio samples*, and *transcripts* pass through ten frozen foundation-model backbones; the fused representation drives a generative model that synthesises whole-brain fMRI activity over the Schaefer-1000 parcellation. Rather than a single point estimate, MultiFlow models the stimulus-conditioned BOLD distribution via Stimulus-Informed Source Flow Matching (SISFM).

over time (Figure 1). Encoding models that predict cortical responses from such stimuli give a principled way to probe this integration [63], and the Algonauts 2025 dataset [0, 8], with tens of hours of whole-brain recordings per participant, makes it feasible to train deep encoders on them.

Concretely, the task is the following. A participant lies in an MRI scanner watching an episode of a television series. Every 1.49 s the scanner records one whole-brain volume, summarised here as 1000 cortical parcels. An *encoding model* sees only what the participant saw and heard, that is the video frames, the soundtrack and the dialogue, and must predict what the scanner recorded, parcel by parcel and timepoint by timepoint; accuracy is the correlation between predicted and measured signal in each parcel. A model that does this well is a testable account of how sensory input drives cortical activity, and lets stimuli or hypotheses be screened without acquiring new scans.

Leading methods on this dataset (TRIBE [6], VIBE [29] and MedARC [60]) establish multimodal fusion as essential for high-fidelity encoding, yet all treat encoding as deterministic regression to a single BOLD point estimate. This discards a defining property of cortical responses: identical stimuli evoke a *distribution* of neural responses. Squared-error optimisation collapses that distribution onto the conditional mean, a collapse hypothesised to underlie the limited out-of-distribution (OOD) generalisation of current encoders [0].

Encoding is therefore recast as a *conditional generative* problem using flow matching [13, 16, 30] (Figure 2). Standard CFM inherits a Gaussian source $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ from image synthesis, where no stimulus exists at inference; in encoding the stimulus is observed at *both* training and inference, yet a Gaussian source ignores it. This work introduces **Stimulus-Informed Source Flow Matching (SISFM)**, which conditions the source on the stimulus: a depthwise temporal convolution shapes the source mean with a receptive field matching the hemodynamic response function (HRF) length, while a Softplus variance head shrinks its scale, shortening the expected transport path. SISFM subsumes vanilla CFM as the special case $\mu_\phi \equiv \mathbf{0}$, $\sigma_\phi \equiv 1$.

SISFM is instantiated as **MultiFlow**: a heterogeneous fusion module over ten frozen vision, audio, language, and audio-visual foundation models [0, 3, 12, 17, 19, 24, 32, 35], a DiT-X decoder [21, 34] parameterising the velocity field with sub-layer Attention Residuals [13] in place of fixed identity skips, and per-subject output heads partitioned by the seven Yeo functional networks [36] of the Schaefer-1000 parcellation [24].

The primary contributions are:

- *SISFM*: a generalisation of CFM that conditions the source distribution on the stimulus through an HRF-length depthwise convolution and a Softplus variance head, trained jointly with the velocity field via a three-term objective; to our knowledge, this is the first flow-matching framework for fMRI encoding.
- The *MultiFlow architecture*: a DiT-X decoder with sub-layer Attention Residuals [14], heterogeneous multi-token fusion over ten frozen backbones, and per-subject Yeonetwork heads.
- An *empirical evaluation* on Algonauts 2025: a single MultiFlow checkpoint trails the 1000-model TRIBE ensemble in-distribution by 0.003 PCC and the strongest OOD ensemble by 0.009 PCC without model averaging and without claiming parity (Tables 1–2), and also delivers stimulus-conditioned uncertainty and sample-based diversity.

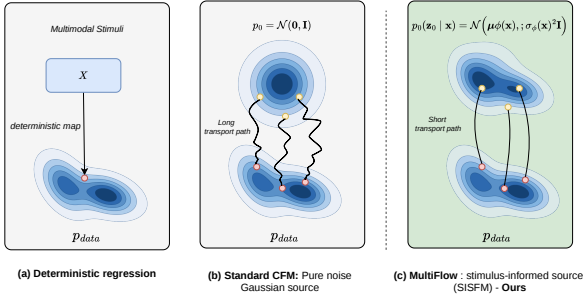


Figure 2: **Three views of fMRI encoding (schematic).** (a) Deterministic regression maps the stimulus $\mathbf{x}$ to a single point estimate $\hat{\mathbf{y}}$ (red), collapsing the target BOLD distribution onto its conditional mean. (b) Standard CFM draws source samples $\mathbf{z}_0$ (orange) from a fixed Gaussian $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and transports each to $p_{\text{data}} \approx p(\mathbf{y} | \mathbf{x})$ along a long ODE path. (c) **MultiFlow (ours)** draws the *same* samples from a *stimulus-conditioned* source $p_0(\mathbf{z}_0 | \mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2 \mathbf{I})$ whose μ_ϕ and σ_ϕ are *learned* to place the source closer to the target, shortening the expected transport path. In (b,c) the curves are independent samples from a *single* model; the panels differ only in the source distribution.

2 Related Work

fMRI encoding from naturalistic stimuli. Classical encoding models map stimulus features to voxel-wise BOLD responses via ridge regression [10, 11, 13]. The three leading entries on Algonauts 2025 share a common recipe: extract time-aligned features from frozen video, audio, and language foundation models, fuse them, and regress the result onto parcel responses with a transformer. They differ in fusion, readout, and ensembling. TRIBE [5] mean-pools per-modality features before fusion, reads out through a single shared per-subject linear head, and averages 1000 runs. VIBE [25] uses a similar frozen-feature pipeline and averages 20 runs; both rely on hard gradient-stopping to keep auxiliary objectives from dominating the encoder. MedARC [16] averages 50 runs and is the strongest published OOD ensemble. All three treat encoding as deterministic regression to a single

BOLD point estimate. This discards BOLD variability arising from physiological noise [8], hemodynamic dispersion [9], and cognitive variability, limiting out-of-distribution (OOD) generalisation [10]. MultiFlow recasts encoding as a conditional generative task to model this full response distribution.

Generative models of neural data. Generative models (e.g., diffusion) are widely used for fMRI *decoding* (brain-to-stimulus) [20, 26, 27]. Elsewhere in medical imaging a generative branch often serves as an auxiliary regulariser rather than as the primary output, as in the VAE-regularised CNN-Transformer of SegTransVAE [22]; we observe a related effect from the SISFM objective (Section 4.6). Encoding (stimulus-to-brain), however, remains under-explored generatively. MultiFlow is a conditional flow-matching (CFM) encoder, modelling the stimulus-conditioned BOLD distribution $p(\mathbf{y} \mid \mathbf{x})$.

Conditional Flow Matching (CFM) and SISFM. CFM [15, 16, 80] learns a velocity field $v_\theta(\mathbf{z}_t, t, \mathbf{x})$ that generates data $\mathbf{y}$ from a source $\mathbf{z}_0 \sim p_0$ via the ODE $\dot{\mathbf{z}}_t = v_\theta$. Under the optimal-transport coupling, the paths are defined by the linear interpolation:

$$\mathbf{z}_t = t\mathbf{y} + (1-t)\mathbf{z}_0, \quad u_t = \mathbf{y} - \mathbf{z}_0. \quad (1)$$

The model is trained by regressing v_θ onto the target velocity u_t :

$$\mathcal{L}_{\text{flow}}(\theta; \phi) = \mathbb{E}_{t, \mathbf{x}, \mathbf{y}, \mathbf{z}_0 \sim p_0(\cdot \mid \mathbf{x})} \|v_\theta(\mathbf{z}_t, t, \mathbf{x}) - (\mathbf{y} - \mathbf{z}_0)\|^2. \quad (2)$$

Standard CFM fixes the source to a stimulus-blind Gaussian $p_0 = \mathcal{N}(\mathbf{0}, \mathbf{I})$, which in encoding ignores an input that is available at inference, increasing the expected transport distance $\mathbb{E}\|\mathbf{y} - \mathbf{z}_0\|^2$ and causing trajectories to cross. We introduce **Stimulus-Informed Source Flow Matching (SISFM)**, which conditions the source on the stimulus:

$$p_0(\mathbf{z}_0 \mid \mathbf{x}) = \mathcal{N}(\mu_\phi(\mathbf{x}), \sigma_\phi(\mathbf{x})^2 \mathbf{I}), \quad (3)$$

where the mean μ_ϕ and variance scale σ_ϕ are learnable: μ_ϕ acts as a shape prior and σ_ϕ shrinks the source ball, leaving the velocity field to model residual variability. Section 3.3 gives the parameterisation.

3 Method

3.1 Problem Formulation and Notation

We now state the task formally. A subject watches a movie sampled at $\text{TR} \approx 1.49$ s. For each window, the input is a multimodal stimulus context $\mathbf{x} \in \mathbb{R}^{T_{\text{ctx}} \times D}$ ($T_{\text{ctx}} = 101$ TRs, $D = 31,872$ features from ten frozen backbones). The target is the corresponding cortical response $\mathbf{y} \in \mathbb{R}^{T_{\text{tgt}} \times V}$ ($T_{\text{tgt}} = 50$ TRs, $V = 1000$ Schaefer-1000 parcels). The context window is centred on the target with symmetric padding (≈ 37 s on each side) to provide narrative history and brief look-ahead. Rather than using deterministic regression to predict a single BOLD point estimate $\hat{\mathbf{y}} = f_\theta(\mathbf{x})$, MultiFlow models the BOLD distribution $p(\mathbf{y} \mid \mathbf{x})$ via flow matching, starting from the stimulus-conditioned source distribution in Eq. (3).

3.2 Heterogeneous Multimodal Fusion

Multimodal fusion. The ten backbones span nearly an order of magnitude in dimensionality ($d = 1,280$ to $6,144$) (Figure 3). Each TR-aligned stream $\mathbf{f}^{(i)}$ is projected to a common

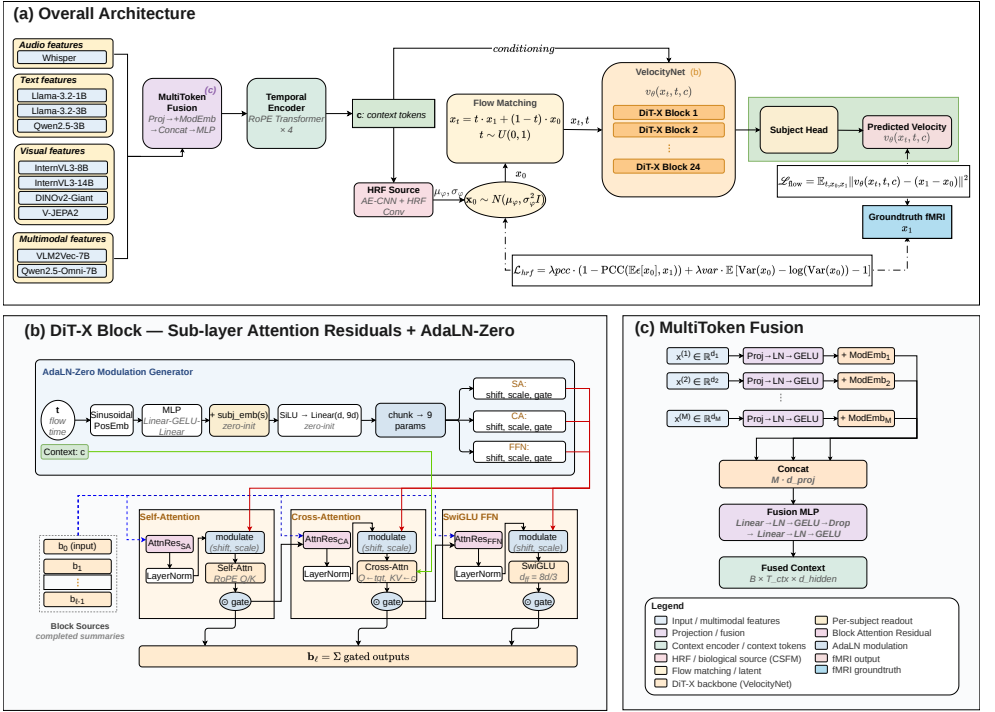


Figure 3: **MultiFlow overview.** Ten frozen foundation backbones feed a multi-token fusion module and a RoPE temporal encoder. The encoded context conditions both the SISFM source generator (producing a stimulus-aligned starting distribution) and the DiT-X velocity decoder. Per-subject Yeo-network heads project the latent state to a 1000-parcel BOLD prediction.

dimension $p = 512$, summed with a learnable modality embedding $\mathbf{e}^{(i)} \in \mathbb{R}^{512}$, and concatenated along the feature axis. A two-layer MLP ($5120 \rightarrow 2048 \rightarrow 1024$) then produces the fused context $\mathbf{c} \in \mathbb{R}^{T_{\text{ctx}} \times 1024}$. Unlike the mean-pooled fusion in TRIBE [5], this preserves per-modality information for later learned mixing.

Temporal encoder. A four-layer pre-norm Transformer with RoPE [49] encodes relative stimulus-to-BOLD lag rather than absolute clip position. It runs bidirectionally over $T_{\text{ctx}} = 101$ TRs and its output is centre-sliced to $T_{\text{tgt}} = 50$ TRs, yielding $\mathbf{c}_{\text{enc}} \in \mathbb{R}^{50 \times 1024}$ shared by the source generator and velocity decoder.

3.3 Stimulus-Informed Source Flow Matching

Source parameterisation. The stimulus-conditioned source distribution $p_0(\mathbf{z}_0 | \mathbf{x})$ in Eq. (3) is parameterised by a temporal operator whose receptive field matches the canonical HRF support window (Figure 4). The mean μ_ϕ uses a Conv1D cascade over $\mathbf{c}_{\text{enc}}$ to produce a non-negative driver, convolved by a depthwise Conv1D whose kernel is unconstrained and whose length matches the canonical HRF support window; non-negativity applies to the driver, not the kernel, so the operator imposes an HRF *support length* rather than an HRF *shape*. The latent mean $\mu_\phi^{\text{lat}} \in \mathbb{R}^{50 \times 1024}$ is projected to voxel space by subject heads (Section 3.5). The

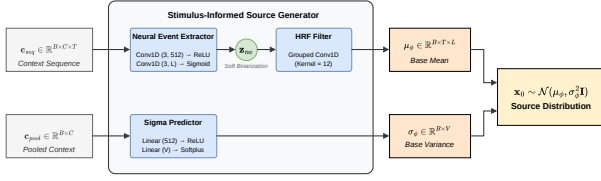


Figure 4: **SISFM source generator.** A Conv1D–ReLU–Conv1D–Sigmoid stage produces a non-negative driver, convolved by a depthwise Conv1D with HRF-length kernel to yield μ_ϕ . A pooled MLP–Softplus head produces σ_ϕ .

variance scale $\sigma_\phi(\mathbf{x}) = \text{softplus}(h_\phi(\mathbf{x})) + 10^{-4}$ is predicted via a temporally pooled context and a two-layer MLP, yielding a per-voxel vector broadcast across TRs.

Training objective. Joint training of (μ_ϕ, σ_ϕ) and the velocity field v_θ is augmented to prevent the degenerate minimiser $(\mu_\phi \rightarrow \mathbf{y}, \sigma_\phi \rightarrow 0)$ under which generative capacity collapses:

$$\mathcal{L}_{\text{SISFM}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{var}} \mathcal{L}_{\text{var}} + \lambda_{\text{pcc}} \mathcal{L}_{\text{pcc}}, \quad (4)$$

$$\mathcal{L}_{\text{var}} = \mathbb{E}[\sigma_\phi^2 - \log \sigma_\phi^2 - 1], \quad \mathcal{L}_{\text{pcc}} = \mathbb{E}[1 - \text{PCC}(\mu_\phi(\mathbf{x}), \mathbf{y})]. \quad (5)$$

$\mathcal{L}_{\text{var}}$ is proportional to the KL divergence between $\mathcal{N}(0, \sigma_\phi^2)$ and a unit Gaussian, penalising variance collapse by pulling σ_ϕ toward unity. $\mathcal{L}_{\text{pcc}}$ provides scale-invariant supervision: maximising Pearson correlation pushes μ_ϕ to track the *shape* of $\mathbf{y}$ without competing with v_θ for amplitude. From a validation grid search we fix $\lambda_{\text{var}} = 0.05$ and $\lambda_{\text{pcc}} = 1.0$; this soft balancing avoids the hard gradient-stopping used in [9, 24]. During training the mean voxel variance $\bar{\sigma}_\phi^2$ converges to ≈ 0.19 while the source residual stays well above prediction error, indicating that the source does not collapse.

3.4 DiT-X Decoder with Attention Residuals

While the source distribution provides a stimulus-aligned starting estimate, the velocity field transforms this state into the target BOLD response. This velocity field is parameterised with a 24-block DiT-X decoder [64] (1024-dim, 16 heads). Each block combines self-attention over target tokens, cross-attention to $\mathbf{c}_{\text{enc}}$ (both with RoPE), and a SwiGLU FFN [28], modulated by AdaLN-Zero [71] with zero-initialised gates.

Sub-layer Attention Residuals. Following [13], the identity skip is replaced with learned softmax-weighted aggregation over all previous block summaries,

$$\text{AR}(\{\mathbf{s}_n\}_{n=0}^N) = \sum_{n=0}^N \alpha_n \mathbf{s}_n, \quad \alpha = \text{softmax}(\mathbf{q}^\top \text{RMSNorm}(\mathbf{s}_n)), \quad (6)$$

where $\mathbf{q} \in \mathbb{R}^{1024}$ is a zero-initialised pseudo-query and summaries include the input embedding and all preceding block outputs. The operator is inserted at all three sub-layer positions, letting each velocity component draw on the depth where its relevant feature is most legible.

3.5 Per-Subject Yeo-Network Heads

The DiT-X output $\mathbf{h} \in \mathbb{R}^{50 \times 1024}$ is mapped to fMRI space by per-subject Yeo-network heads. For each of the seven Yeo networks [66] (Visual, SomMot, DorsAttn, SalVentAttn, Limbic,

Cont, Default) the head holds a weight tensor $W^{(k)} \in \mathbb{R}^{N_s \times 1024 \times 2n_k}$, with $N_s = 4$ subjects and $2n_k$ the bilateral parcel count for network k (so $\sum_k 2n_k = V = 1000$). At inference the relevant subject’s slice is applied to $\mathbf{h}$ and the seven outputs are concatenated to give $\hat{\mathbf{y}} \in \mathbb{R}^{50 \times 1000}$. This lets functionally distinct cortical systems specialise their readout while sharing the up-stream representation, a finer grain of subject conditioning than the single shared linear head used in TRIBE [6].

3.6 Inference

At test time, K independent samples $\mathbf{z}_0^{(k)} \sim p_0(\cdot | \mathbf{x})$ are drawn from the SISFM source (with temperature $\tau=0.2$ scaling σ_θ) and the probability-flow ODE is integrated for each through the *same* learned velocity field with a midpoint solver for $N = 30$ steps on a square-root time grid $t_i = \sqrt{s_i}$, which places finer steps near $t = 1$ where the trajectory becomes sharper. The K predictions are then averaged, and long clips use sliding-window inference with overlap averaging. Unless stated otherwise, all reported numbers use $K = 5$, at which the sampling gain saturates (Tables 1 and 2). We call this *source-averaged inference*, to make explicit that the K draws come from the SISFM source of a *single* trained checkpoint, unlike baselines that average 20–1000 re-trained models.

4 Experiments

4.1 Setup

We evaluate on the Algonauts 2025 / CNeuroMod dataset [4, 8] using subjects sub-01, 02, 03, and 05 (sub-04 is excluded due to incomplete data). Training uses Friends S1–5 and four feature movies; S6 is validation, S7 is the in-distribution (ID) test split, and six held-out movies are the out-of-distribution (OOD) test split. fMRI volumes are z-scored, and we report per-parcel Pearson correlation (PCC) over the Schaefer-1000 parcellation. Ten frozen V/A/L/AV backbones are pooled to TR resolution ($D = 31,872$; per-backbone details in the supplementary material). Training runs on a single RTX 3090 GPU for 100 epochs with AdamW (lr = 10^{-4} , cosine decay, 5% warm-up), bfloat16, effective batch size 512, and weight EMA (decay 0.999); windows use $T_{\text{ctx}} = 101$, $T_{\text{tgt}} = 50$, stride 10, and jitter ± 2 TRs.

Baselines and protocol. We compare MultiFlow (one trained model) against: (i) published ensemble results of TRIBE [6] (1000 models), VIBE [25] (20 models), and MedARC [60] (50 models, OOD only); (ii) reproduced single-run baselines (one trained model) under matched features/splits; and (iii) *Regression DiT-X*, a deterministic control. Main results and ablations use DiT-X-Large (24 blocks, 1024-dim, 16 heads) trained on all subjects; ablations report mean S6 PCC.

All single-run baselines are retrained from the authors’ published code under our features, data, and splits, with their released hyperparameters. One caveat applies to the reproduced TRIBE: its mean-pooling fusion was designed for the three-backbone set it was built around, so under our ten-backbone protocol it pools a considerably more heterogeneous input, which likely explains why it scores lowest of the single-run baselines despite being the strongest published ensemble. We therefore do not treat it as an upper bound on that method, and read the single-run column as a comparison of fusion designs under *our* protocol rather than a ranking of the original systems.

Method	#Trained models	#Source draws	Sub-01	Sub-02	Sub-03	Sub-05	Mean
Ridge [8]	–	–	0.2152	0.2031	0.2189	0.1759	0.2033
<i>Single-run, matched training compute (1 trained model, same features/data/splits)</i>							
TRIBE [9]	1	–	0.2827	0.2689	0.2817	0.2344	0.2670
VIBE [14]	1	–	0.3201	0.3042	0.3093	0.2595	0.2983
MedARC [16]	1	–	0.3271	0.3097	0.3209	0.2642	0.3055
Regression DiT-X (ours, det.)	1	–	0.2887	0.2734	0.2807	0.2331	0.2690
<i>Full ensemble baselines</i>							
TRIBE [9]	1000	–	0.3419	0.3239	0.3346	0.2775	0.3195
VIBE [14]	20	–	0.3503	0.3309	0.3410	0.2849	0.3268
MultiFlow (ours)	1	5	0.3398	0.3217	0.3275	0.2760	0.3163

Table 1: **In-distribution performance (Friends S7).** Per-subject and mean per-parcel PCC ($\uparrow$) over Schaefer-1000. *#Trained models* counts independently trained checkpoints; *#Source draws* counts samples averaged from the SISFM source of a *single* checkpoint (source-averaged inference, Section 3.6), which adds no training cost. “Matched training compute” refers to the *training* budget: all single-run rows share MultiFlow’s features, data, splits, and frozen backbones, whose inference cost dominates. *Regression DiT-X* is identical to MultiFlow except that the SISFM source and variance head are removed and MSE is used, so its gap to MultiFlow isolates the generative formulation. MultiFlow uses $K = 5$ draws ($\tau = 0.2$); bold is best.

4.2 Main Results

Analysis of main results. Results are analysed in-distribution (ID, Table 1) and out-of-distribution (OOD, Table 2). Two controls attribute any improvement to the right cause: a *matched-training-compute* control (all single-run rows share MultiFlow’s features, data, and splits) and a *matched-architecture* control (*Regression DiT-X*, which removes the SISFM source but keeps the backbone fixed and is trained with MSE).

System level (matched compute). Among single-run models, MultiFlow attains the highest mean PCC on both splits at $K=5$ source draws. Specifically, on ID, it scores 0.3163 vs. 0.2670 (TRIBE), 0.2983 (VIBE), and 0.3055 (MedARC); on OOD, it scores 0.2080, above all three single-run baselines (TRIBE 0.1705, VIBE 0.1968, MedARC 0.2054). At a single draw ($K=1$) the margins narrow and the OOD ordering no longer holds (0.1983 against MedARC’s 0.2054). This makes the full system strongest at $1\times$ training compute, but because these baselines differ in *both* architecture and objective, it does not by itself attribute the gain to the generative formulation.

Formulation level (matched architecture). Holding the DiT-X backbone fixed and replacing MSE with SISFM lifts mean PCC from 0.2690 to 0.3163 on ID and from 0.1318 to 0.2080 on OOD (*Regression DiT-X* vs. MultiFlow, Tables 1–2), isolating the contribution of the generative formulation on the held-out test splits. The architecture alone does not exceed the strongest single-run baseline: *Regression DiT-X* (0.2690) falls 0.0365 below MedARC (0.3055) on ID, and SISFM closes that deficit and adds +0.0108, giving the full +0.0473 gain over the architecture-matched control. On OOD the effect is larger: SISFM contributes +0.0762, lifting the architecture from 0.1318, below the 0.1705–0.2054 range of

Method	#Trained models	#Source draws	Sub-01	Sub-02	Sub-03	Sub-05	Mean
Ridge [8]	–	–	0.0986	0.0859	0.1021	0.0715	0.0895
<i>Single-run, matched training compute (1 trained model, same features/data/splits)</i>							
TRIBE [6]	1	–	0.1875	0.1726	0.1838	0.1381	0.1705
VIBE [29]	1	–	0.2172	0.1921	0.2135	0.1646	0.1968
MedARC [60]	1	–	0.2282	0.1958	0.2275	0.1702	0.2054
Regression DiT-X (ours, det.)	1	–	0.1372	0.1291	0.1463	0.1147	0.1318
<i>Full ensemble baselines</i>							
TRIBE [6]	1000	–	0.2381	0.2105	0.2377	0.1720	0.2146
VIBE [29]	20	–	0.2392	0.2080	0.2294	0.1721	0.2122
MedARC [60]	50	–	0.2458	0.2071	0.2384	0.1772	0.2171
MultiFlow (ours)	1	5	0.2319	0.2033	0.2218	0.1752	0.2080

Table 2: **Out-of-distribution performance (6 held-out movies).** Per-subject and mean per-parcel PCC ($\uparrow$) over Schaefer-1000. Columns as in Table 1: *#Trained models* counts independently trained checkpoints, *#Source draws* counts averaged samples from a single checkpoint. *Regression DiT-X* is the architecture-matched MSE control (see Table 1). MultiFlow uses $K = 5$ draws ($\tau = 0.2$). Bold denotes the best result.

the single-run baselines, to 0.2080 above all of them; Section 4.6 attributes this to an overfitting regularisation effect. This reproduces on S7/OOD the same-architecture trend seen on S6 in Table 4 (MSE 0.2760 $\rightarrow$ CFM 0.2973 $\rightarrow$ SISFM 0.3213).

Comparison with ensembles. Against published ensemble results, single-checkpoint MultiFlow trails the 1000-model TRIBE ensemble on ID by only 0.003 PCC (0.3163 vs. 0.3195), trailing the strongest ID ensemble (VIBE-20) by 0.010 PCC. On OOD, MultiFlow trails the strongest ensemble (MedARC-50) by 0.009 PCC (0.004–0.009 across ensembles), while on Sub-05 it approaches the best ensemble (0.1752 vs. MedARC-50’s 0.1772). We do not claim parity with ensembled SOTA on accuracy; rather, MultiFlow recovers most of the ensembling benefit using a single trained model rather than an ensemble of 20–1000 independently retrained checkpoints. A significance test on these margins is not possible: S7 and the OOD movies are scored by the provider’s private server, which returns only an aggregate PCC, and the published ensembles release no per-parcel outputs. We therefore report the ensemble margins descriptively and disclaim parity; Section 4.6 gives per-parcel statistics for the testable comparisons, those against our own architecture-matched control.

Source-averaging gain. Drawing $K = 5$ samples from one checkpoint improves over a single sample ($K = 1$): 0.3057 $\rightarrow$ 0.3163 on ID and 0.1983 $\rightarrow$ 0.2080 on OOD, after which the gain saturates.

Generative capabilities. Beyond accuracy, sampling from the SISFM source yields samples from the predictive distribution and a stimulus-conditioned scale $\sigma_\phi(\mathbf{x})$ from the same checkpoint, which are capabilities a single deterministic regressor does not provide.

4.3 Cross-Subject Generalisation

Generalisation to unseen subjects probes how much of a model’s fit is carried by subject-specific readout parameters rather than by the shared upstream representation. Since TRIBE,

VIBE, and MedARC do not report cross-subject results on this dataset, evaluation is performed using reproduced single-run baselines (same features, pipeline, no ensembling). This ensures a matched-training-compute comparison that isolates algorithmic differences from ensembling effects. Ensembled cross-subject numbers are unavailable for comparison: none are published, and producing them would require retraining each baseline’s full 20–1000-model ensemble once per leave-one-subject-out fold, so Table 3 is a single-run comparison by necessity. Table 3 reports leave-one-subject-out PCC: models trained on three subjects, evaluated on the held-out fourth, averaged over three clips (*Friends S6*, *Figures*, *Life*). For the held-out subject, inference uses the mean of the trained subjects’ Yeo-network heads as the readout projection, so no subject-specific parameters are available for the test subject.

Method	Hold-out (test) subject (model trained on remaining 3)			
	Sub-01 train: 02, 03, 05	Sub-02 train: 01, 03, 05	Sub-03 train: 01, 02, 05	Sub-05 train: 01, 02, 03
TRIBE [5]	0.1625	0.1584	0.1691	0.1242
VIBE [24]	0.2376	0.2442	0.2499	0.2336
MedARC [6]	0.2007	0.2040	0.2232	0.1855
MultiFlow (ours)	0.2388	0.2507	0.2502	0.2442

Table 3: **Cross-subject generalisation (leave-one-subject-out)**. Column indicates the held-out test subject. PCC is averaged over three clips (*Friends S6*, *Figures*, *Life*). Baselines are single-run reproductions (no ensembling, same features/pipeline as MultiFlow). Bold denotes the best result.

Analysis. Under matched single-run conditions, MultiFlow matches or exceeds the strongest single-run baseline (VIBE [24]) on every held-out subject, with the largest margin on Sub-05 (0.2442 vs. 0.2336). We attribute this to two factors: the generative formulation regularises implicitly (Section 4.6), reducing overfitting to training subjects, and the per-subject Yeo-network heads share an upstream representation, which eases transfer when they are averaged for an unseen subject. The absolute numbers set the interpretation: leave-one-subject-out PCC spans 0.239–0.251 against 0.3163 in-distribution, a shortfall of roughly a quarter of the ID score. Table 3 should therefore be read as evidence of *relative robustness* under a matched protocol, meaning MultiFlow degrades less than the baselines when the test subject is unseen, and not as evidence of a subject-invariant representation. Establishing the latter would require broader subject coverage and readouts that are meta-learned rather than averaged (Section 5).

4.4 Spatial Mapping and Qualitative Visualisation

We project per-parcel S6 PCC onto the Schaefer-1000 inflated cortical surface and, for six representative regions, plot region-averaged predicted BOLD against ground truth over a full episode.

Cortical hierarchy and temporal fidelity. Figure 5 (top) shows the classical encoding gradient [33]: per-parcel PCC peaks in ventral and lateral occipito-temporal cortex (V4, fusiform, STS, TPOJ), decays anteriorly, and varies smoothly across the surface. The traces (bottom) make the hierarchy concrete: TPOJ ($R=0.58$), STS / FFA ($R=0.54$) and V4 ($R=0.53$) sit in the upper achievable range, while PEF ($R=0.49$) and the small-n medial V2 cluster ($R=0.41$) drop slightly but stay temporally aligned. For $R \geq 0.5$ regions the prediction tracks

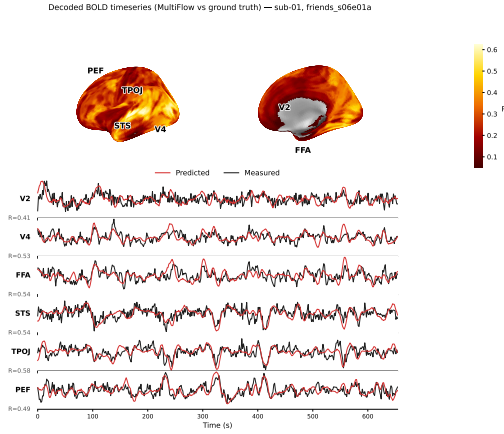


Figure 5: **Predicted BOLD timeseries across the cortical hierarchy (sub-01, Friends s06e01a).** *Top:* per-parcel PCC on the fsaverage5 inflated surface (LH lateral, RH medial), six regions annotated. *Bottom:* predicted (red) vs. measured (black) BOLD averaged over each region’s parcels across ~ 656 s; parcel-mean R on the full S6 holdout is given next to each track.

BOLD peaks and troughs at the TR scale, including excursions above 1 SD lasting 10–30 s, consistent with the HRF-length source mean capturing much of the temporal structure before the velocity field acts.

4.5 Component Ablations

All ablations use the DiT-X-Large variant and are trained on all four subjects (sub-01, 02, 03, 05) using Friends Seasons 1–5 plus the four feature movies, with Friends Season 6 held out for evaluation (mean PCC across the four subjects). S6 is the benchmark’s designated validation split and its ground truth is available locally, enabling rapid iterative comparison across many ablation variants. The S7 and OOD labels are withheld by the dataset provider and can be scored only through its evaluation server, which returns a single aggregate PCC per submission; iterative ablation and per-parcel analysis are therefore not possible on those splits, and they are reserved for the final system-level evaluation in Tables 1–2.

Generative versus deterministic encoding. Table 4 compares three formulations of the same architecture: Regression DiT-X (MSE loss), CFM with a standard Gaussian source, and the full SISFM model.

Replacing MSE regression with standard-Gaussian CFM improves PCC by +0.0213, indicating that a generative formulation benefits encoding even without a stimulus-conditioned source. SISFM adds a further +0.0240, the largest single gain: by learning an HRF-length source mean μ_ϕ and variance σ_ϕ , it offloads basic hemodynamics from the velocity field, which then models only residual variability. The total gain over regression is +0.0453 (+16.4% relative).

Decoder backbone. Table 5 compares the DiT-X decoder against an MLP, a UNet, and 1D-DiT without cross-attention.

MLP and UNet decoders perform comparably low (0.2565 vs. 0.2592): MLPs ignore temporal structure, and UNets are constrained by the local receptive field of 1D convolutions.

Model	Source p_0	Training loss	S6 PCC ($\uparrow$)
Regression DiT-X	-	MSE	0.2760
CFM + Gaussian	$\mathcal{N}(\mathbf{0}, \mathbf{I})$	$\mathcal{L}_{\text{flow}}$	0.2973
CFM + SISFM	$p_0(\cdot \mathbf{x})$	Eq. 4	0.3213

Table 4: **Generative vs. deterministic encoding.** Same architecture, three formulations. Across three inference seeds the MSE→CFM→SISFM ordering is preserved, with a per-row range ≤ 0.002 PCC, well below the reported gaps.

Decoder	S6 PCC ($\uparrow$)
MLP (3-layer)	0.2565
UNet	0.2592
DiT	0.3035
DiT-X + Attention Residuals	0.3213

Table 5: **Decoder backbone.** Effect of cross-attention and sub-layer Attention Residuals.

Replacing the UNet with a self-attentive DiT adds +0.0443, indicating that BOLD at any TR depends on long-range temporal context. Adding cross-attention and sub-layer Attention Residuals (DiT-X + AR) contributes a further +0.0178: cross-attention in every block lets the decoder query the stimulus context at different temporal lags, and Attention Residuals allow depth-selective routing between low-level sensory and semantic features.

Modality coverage. Table 6 isolates the contribution of each modality group through a factorial ablation.

Fusing several backbones already helps within a modality: the strongest *single* visual backbone (InternVL3-14B, 0.1932) trails the fused V-only setting (0.2043), so part of the gain is within-modality complementarity rather than cross-modal binding. Language alone (0.217) slightly outperforms vision alone (0.204), consistent with the recruitment of semantic and social networks during narrative viewing [14]. Pairwise combinations exceed their constituents, V+A (0.282) over V-only (0.204) and A-only (0.159), and adding jointly pre-trained audio-visual backbones to V+A+L gives the largest marginal gain, from 0.3020 to 0.3213, consistent with joint pre-training capturing cross-modal binding.

4.6 Generative Analysis: Robustness, UQ, and Diversity

This section evaluates step robustness, uncertainty quantification (UQ), and sample diversity on the validation split (Friends S6).

Path length and step-count robustness. Straightness of the learned velocity field is governed by the source–data coupling, whose expected squared displacement is $\mathbb{E}_{\epsilon} \|\mathbf{y} - \mathbf{z}_0\|^2$ (with $\mathbf{z}_0 = \mu_{\phi} + \sigma_{\phi}\epsilon$). This equals $\|\mathbf{y}\|^2 + d$ for a Gaussian source and $\|\mathbf{y} - \mu_{\phi}\|^2 + d\bar{\sigma}_{\phi}^2$ for SISFM ($d = 50,000$, with $\bar{\sigma}_{\phi}^2 \approx 0.19$ the converged training-time variance). On S6 the expected length drops from 98.3k to 57.4k (−41.6%), almost entirely through variance shrinkage: the noise term $d\bar{\sigma}_{\phi}^2$ contracts from 50k to 9.5k, while the mean term $\|\mathbf{y} - \mu_{\phi}\|^2$ falls by less than 1% (from $\sim 48.3\text{k}$ to $\sim 47.9\text{k}$), because the scale-invariant objective $\mathcal{L}_{\text{pcc}}$ makes μ_{ϕ} a shape-aligned prior (PCC 0.26) rather than an amplitude-matched one. The tighter

Setting	V	A	L	AV	S6 PCC $\uparrow$
V only	✓	-	-	-	0.2043
A only	-	✓	-	-	0.1597
L only	-	-	✓	-	0.2170
V + A	✓	✓	-	-	0.2821
V + L	✓	-	✓	-	0.2995
A + L	-	✓	✓	-	0.2872
V + A + L	✓	✓	✓	-	0.3020
All	✓	✓	✓	✓	0.3213

Table 6: **Modality coverage.** Factorial ablation over the V, A, L, and AV groups: single-modality rows give standalone predictive power, pairwise rows show cross-modal complementarity, and the last two rows give the marginal gain of the full stack.

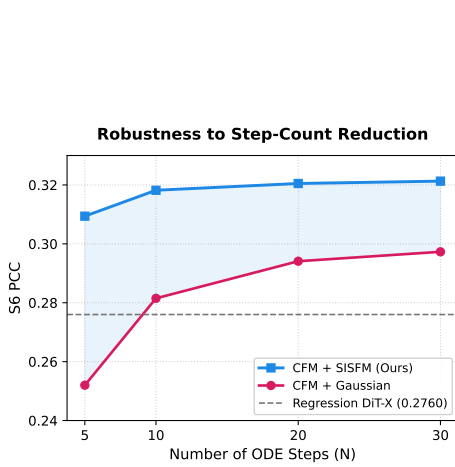


Figure 6: Step-count robustness (S6 PCC). From $N=30$ to $N=5$ steps, SISFM (blue) retains accuracy while standard CFM (red) decays, falling below the Regression DiT-X baseline (dashed grey) at low N .

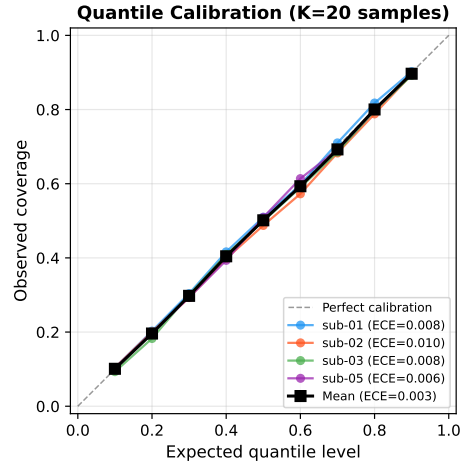


Figure 7: Quantile calibration (S6, $K=20$). Observed coverage vs. expected quantile level; proximity to the diagonal indicates calibrated predictive intervals from a single checkpoint. Per-subject expected calibration error ranges over 0.006–0.010.

coupling yields straighter trajectories, and at inference samples are drawn closer still to μ_ϕ (temperature $\tau=0.2$ scaling σ_ϕ), so the probability-flow ODE integrates accurately under few steps. As shown in Table 7 and Figure 6, reducing ODE steps from $N=30$ to $N=5$ collapses standard CFM to 0.2520 while SISFM holds 0.3094, widening Δ from $+0.0240$ to $+0.0574$ and confirming that a shorter transport path buys step-efficiency at test time.

ODE steps N	5	10	20	30
CFM + Gaussian	0.2520	0.2815	0.2941	0.2973
CFM + SISFM	0.3094	0.3182	0.3205	0.3213
Δ (SISFM–Gauss)	+0.0574	+0.0367	+0.0264	+0.0240

Table 7: Step-count robustness (S6 PCC, $\uparrow$). The SISFM source preserves accuracy under step-count reduction better than a Gaussian source, the behaviour predicted by its shorter expected transport path.

Source averaging and regularisation. At test time, drawing $K=5$ independent samples $\mathbf{z}_0^{(k)} \sim p_0$ and integrating them through the same velocity field (source-averaged inference) improves performance from 0.3057 ($K=1$) to 0.3163 ($K=5$) on ID, recovering the variance reduction of deep ensembles without retraining (Tables 1–2). The generative formulation also acts as a regulariser: under MSE with identical hyperparameters, Regression DiT-X drops 51.0% in mean PCC from ID to OOD ($0.2690 \rightarrow 0.1318$), whereas MultiFlow drops 34.2% ($0.3163 \rightarrow 0.2080$), comparable to the 32–34% of VIBE ($0.2983 \rightarrow 0.1968$) and MedARC ($0.3055 \rightarrow 0.2054$) while remaining more accurate on both splits.

Uncertainty quantification and calibration. We run inference with $K=20$ samples on S6. First, the learned per-voxel predictive scale $\sigma_\phi(\mathbf{x}) \in \mathbb{R}^V$ correlates with per-voxel prediction

error (Spearman $\rho = 0.36 \pm 0.04$, $p < 10^{-5}$), so the source variance is a usable confidence signal. Second, treating the spread of the $K = 20$ samples as a predictive distribution, the expected calibration error, the mean absolute gap between observed coverage and expected quantile level over 9 levels ($p \in \{0.1, \dots, 0.9\}$), ranges from 0.006 to 0.010 across the four subjects (Figure 7). Third, the sample standard deviation does not collapse (mean 0.32 ± 0.03) and is higher where error is larger (Spearman $\rho = 0.42$, $p < 10^{-6}$), so MultiFlow expresses heteroscedastic uncertainty rather than degenerate point estimates.

Scope and biological interpretability. Three qualifications bound these results. First, the calibration result is a property of the *transported sample distribution*, not of the magnitude of σ_ϕ : the source scale is rescaled at inference by $\tau=0.2$, so calibration is produced by the source and the velocity field jointly, and σ_ϕ alone carries no calibrated interval. Second, σ_ϕ is stimulus-conditioned *by construction*, being a function of $\mathbf{x}$; what we demonstrate is that its spatial profile tracks per-voxel error, not that it varies with stimulus *content*. Third, our experiments cannot say how much of the modelled variability is genuine neurophysiological noise and how much is epistemic, reflecting what the model has not learned from limited data. That separation requires an observable this dataset lacks under a held-out protocol: repeated presentations of the same stimulus to the same participant, or a parameter ensemble whose disagreement isolates the epistemic component. We therefore retain two readings of the SISFM source, as an implicit regulariser (the ID $\rightarrow$ OOD degradation above) and as an empirically calibrated predictive distribution, and retire a third: we do not claim σ_ϕ measures biological response variability. A repeat-presentation protocol, or a parameter ensemble evaluated against measured test–retest variance, would settle this.

5 Conclusion

Limitations and future work. (i) σ_ϕ is a *predictive* scale, and whether it decomposes into physiological and cognitive components is untested here: the dataset offers no repeated presentations under a held-out protocol, so σ_ϕ is validated against prediction error rather than measured response variability (Section 4.6). It is also per-voxel but shared across TRs within a window; a graph-structured or per-TR covariance could capture spatiotemporal heteroscedasticity, and lightweight adaptation of the frozen backbones may benefit the Default Mode and Control networks. (ii) Cross-subject results establish relative robustness, not subject-invariance: leave-one-subject-out PCC (0.239–0.251) sits well below in-distribution accuracy (0.3163), and transfer still relies on per-subject heads and on SISFM hyperparameters selected on `sub-01`. Broader subject coverage and meta-learned readouts are the route forward. (iii) Inference cost constrains online use: a midpoint solver at $N=30$ costs roughly 60 network evaluations per source draw, short of real-time BCI budgets. The shortened transport path buys headroom, since at $N=5$ SISFM still reaches 0.3094 against 0.2760 for the deterministic control and 0.2520 for Gaussian CFM (Table 7), and distillation would extend this further.

Summary. SISFM conditions the flow-matching source on the stimulus; its instantiation MultiFlow reaches 0.3163 ID and 0.2080 OOD PCC from one checkpoint, within 0.003 and 0.009 PCC of 20–1000-model ensembles, and on a fixed architecture lifts S6 PCC from 0.2760 to 0.3213 (Table 4) while also providing a per-voxel predictive scale and calibrated sample spread (Section 4.6). The principle extends wherever a fixed Gaussian source ignores information available at inference, as in speech synthesis [14], molecular docking [9], and molecular design [18].

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT), the ITRC (Information Technology Research Center) support program (IITP-2026-RS-2024-00437718) supervised by IITP.

References

- [1] Mahmoud Assran et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- [2] Julie Boyle, Basile Pinsard, Valentina Borghesani, Francois Paugam, Elizabeth DuPre, and Pierre Bellec. The Courtois NeuroMod project: Quality assessment of the initial data release (2020). In *2023 Conference on Cognitive Computational Neuroscience*, volume 10, pages 2023–1602, 2023.
- [3] Zhe Chen et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.
- [4] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023.
- [5] Stéphane d’Ascoli, Mike Doyle, Hubert Banville, Charlotte Caucheteux, Mohamed El Karoui, and Jean-Rémi King. TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction. *arXiv preprint arXiv:2507.22229*, 2025. 1st place, Algonauts 2025 Challenge. To appear in ICLR 2026.
- [6] Karl J. Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.
- [7] Alessandro T. Gifford, Domenic Bersch, Gemma Roig, and Radoslaw M. Cichy. Insights from the Algonauts 2025 winners. *arXiv preprint arXiv:2508.10784*, 2025.
- [8] Alessandro T. Gifford, Domenic Bersch, Marie St-Laurent, Basile Pinsard, Julie Boyle, Lune Bellec, Aude Oliva, Gemma Roig, and Radoslaw M. Cichy. The Algonauts Project 2025 Challenge: How the human brain makes sense of multimodal movies. *arXiv preprint arXiv:2501.00504*, 2025.
- [9] Gary H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9(4):416–429, 1999.
- [10] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.
- [11] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

- [12] Ziyang Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhui Chen. VLM2Vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*, 2024.
- [13] Kimi Team and Moonshot AI. Attention residuals. *arXiv preprint arXiv:2603.15031*, 2026.
- [14] Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashed Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voice-box: Text-guided multilingual universal speech generation at scale with flow matching. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [15] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.
- [16] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.
- [17] Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [18] Viet Thanh Duy Nguyen, Phuc Pham, and Truong-Son Hy. Enabling multi-target drug discovery through latent evolutionary optimization and synthesis-aware prioritization (EVOSYNTH). *Communications Chemistry*, 9(1):133, 2026. doi: 10.1038/s42004-026-01945-4.
- [19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024.
- [20] Furkan Ozcelik and Rufin VanRullen. Brain-diffuser: Natural scene reconstruction from fMRI signals using generative latent diffusion. *Scientific Reports*, 13, 2023.
- [21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.
- [22] Quan-Dung Pham, Hai Nguyen-Truong, Nam Nguyen Phuong, Khoa N. A. Nguyen, Chanh D. T. Nguyen, Trung Bui, and Steven Q. H. Truong. SegTransVAE: Hybrid CNN-transformer with regularization for medical image segmentation. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. doi: 10.1109/ISBI52829.2022.9761417.
- [23] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, pages 28492–28518, 2023.

- [24] Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex*, 28(9):3095–3114, 2018.
- [25] David Schwetlick et al. VIBE: Video-input brain encoder for fMRI response modeling. *arXiv preprint arXiv:2507.17958*, 2025. 2nd place, Algonauts 2025 Challenge.
- [26] Paul S. Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, Kenneth A. Norman, and Tanishq M. Abraham. Reconstructing the mind’s eye: fMRI-to-image with contrastive learning and diffusion priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [27] Paul S. Scotti, Mihir Tripathy, Cesar Torrico Villanueva, Reese Kneeland, Tong Chen, Ashutosh Narang, Charan Santhirasegaran, Jonathan Xu, Thomas Naselaris, Kenneth A. Norman, and Tanishq M. Abraham. MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In *International Conference on Machine Learning (ICML)*, 2024.
- [28] Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.
- [29] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024.
- [30] Alexander Tong, Nikolay Malkin, Guillaume Huguette, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.
- [31] Cesar Kadir Torrico Villanueva et al. Predicting brain responses to natural movies with multimodal LLMs. *arXiv preprint arXiv:2507.19956*, 2025.
- [32] Jin Xu et al. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.
- [33] Daniel L. K. Yamins and James J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, 2016.
- [34] Ge Yan, Yuquan Deng, Ankit Goyal, Ranjay Krishna, Dieter Fox, Priya Sundareshan, and Ajay Mandlekar. ManiFlow: A general robot manipulation policy via consistency flow training. In *Conference on Robot Learning (CoRL)*, 2025.
- [35] An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- [36] B. T. Thomas Yeo, Fenna M. Krienen, Jorge Sepulcre, Mert R. Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L. Roffman, Jordan W. Smoller, Lilla Zöllei, Jonathan R. Polimeni, Bruce Fischl, Hesheng Liu, and Randy L. Buckner. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165, 2011.