

StimFlow: Synthesizing Brain Activity from Images via Stimulus-conditioned Flow Matching

Bui Van Hung¹, Dang Thanh Vu², Hoang-Son Vo¹, Mai Duc Minh Huy³, Mai Tien Dung⁴, and Soo-Hyung Kim¹

¹ Chonnam National University, Gwangju, South Korea

² AISEED, Inc., South Korea

³ Ho Chi Minh University of Science, Ho Chi Minh City, Vietnam

⁴ MIREA – Russian Technological University, Moscow, Russia

Abstract. The inherent variability of fMRI responses across trials encourages a generative view of the stimulus-to-response mapping. Prior generative encoders often operate in compressed latent spaces and condition predominantly on semantic features, which provide limited low-level cues for early visual cortex. We propose **StimFlow**, which synthesizes fMRI directly in **native voxel space** using **conditional flow matching** with a linear path. This design aligns training and inference distributions, enabling competitive performance with a single sample and reducing reliance on averaging. To match the visual hierarchy, StimFlow combines structural **DINOv2** features with semantic **CLIP** features, which prove complementary along the cortical processing stream. With a single sample, StimFlow achieves a per-voxel Pearson correlation of **0.371**, matching baselines that average five draws. Generation in voxel space allows direct ROI analysis and concept localization without a decoder. Furthermore, it recovers over half of full-data performance from just one hour of scans, supporting efficient cross-subject adaptation.

Keywords: Conditional flow matching · fMRI synthesis · Brain encoding · Native-space modeling · Cross-subject generalization

1 Introduction

Understanding how the human brain encodes natural visual stimuli is a central goal in computational neuroscience and brain–computer interfaces (BCI) [10, 23, 24]. Owing to physiological noise, cognitive-state fluctuations, and hemodynamic variability, the stimulus-to-response mapping is inherently stochastic [1]; even the more modest goal of approximating the *conditional mean* benefits from a generative formulation whose training and inference distributions are aligned, rather than from a deterministic regressor or a diffusion prior whose higher-curvature paths necessitate multi-sample averaging for stability.

Recent generative encoders pursue this goal but share two limitations. First, they synthesize in a *compressed, CLIP-aligned latent space* and are evaluated mainly through a decoded image; because the signal is recovered through an

autoencoder rather than generated in voxel space, fidelity at the voxel and ROI level remains largely unexamined. Second, they condition *exclusively* on language-supervised CLIP features [28], which favour high-level semantics and retain little of the low-level structure (edges, textures, spatial frequency) driving early visual areas (V1–V3), so a single semantic stream is poorly matched to the full functional hierarchy of the visual cortex.

The two representative encoders also differ in how they reach the data distribution: MindSimulator [22] places a diffusion prior over fMRI autoencoder latents and relies on multi-trial averaging and post-selection to obtain stable predictions, whereas SynBrain [2] uses a one-step mapper that is stable but acts as a deterministic predictor. We seek a formulation whose training and inference paths are aligned by construction, so that a single sample ($T=1$) already yields a faithful estimate of the conditional mean without post-hoc averaging.

We propose **StimFlow**, which addresses both axes. StimFlow casts fMRI synthesis as Conditional Flow Matching (CFM) [17,18] along a conditional linear path with independent coupling, generating directly in *native voxel space* with a 1D Diffusion Transformer (DiT-1D) [27]; because the training interpolation starts exactly at the Gaussian distribution inference begins from, training and inference are aligned by construction. To match the cortical hierarchy, it pairs a structural stream (DINOv2 [26]) with a semantic stream (CLIP), which we find functionally complementary: structural features lead in early-to-mid visual areas, semantic features in higher ventral and lateral streams. A shared trunk with lightweight per-subject adapters further enables few-shot adaptation to novel subjects.

On the Natural Scenes Dataset [1], a *single* StimFlow sample reaches a per-voxel Pearson correlation of 0.371, matching or exceeding baselines that average five draws at the voxel level. More importantly, because generation stays in voxel space, fidelity can be read directly per ROI without a separate decoder, and synthesized concept-selectivity maps align closely with the ground truth ($r = 0.94$, top-10% Dice = 0.81), supporting data-driven cortical analysis.

In summary, our contributions are:

- We formulate fMRI synthesis as **conditional flow matching in native voxel space**, enabling direct per-ROI readout and concept localization without a decoder; a single sample matches or exceeds multi-sample baselines on NSD [1] at both the voxel and semantic level.
- We show that **CLIP-only conditioning is suboptimal**: a structural stream (DINOv2) matches or exceeds CLIP alone, and the two are functionally complementary along the cortical hierarchy, motivating dual-stream conditioning.
- We **adapt the DiT architecture to 1D voxel sequences and equip it with per-subject adapters**, supporting few-shot adaptation to novel subjects.

2 Related Work

Brain visual encoding. Visual encoding models predict voxel-level fMRI responses from images and have long been used to probe the functional organization of visual cortex [24]. Early work fit explicit feature spaces (Gabor-wavelet, population receptive-field models) by linear regression [10, 15, 23]; the deep-learning era replaced these with hierarchical CNN activations predicting ventral-stream responses [8, 9, 35], and recent encoders use self-supervised (DINOv2 [26]) or language-supervised (CLIP [28]) features. Yet these encoders remain deterministic, mapping each image to a single response and thus unable to express the trial-to-trial variability intrinsic to fMRI [1].

Brain decoding and reconstruction. The inverse problem (reconstructing the seen image from fMRI) has progressed from early movie reconstruction [25] to diffusion-prior methods [29, 30, 34] that map fMRI into CLIP space and condition an image generator, with MindEye2 enabling strong low-data cross-subject transfer. Decoding is complementary to our goal of synthesizing responses from stimuli; we reuse a frozen MindEye2 [30] only to decode synthesized responses and check their semantic content (Sec. 4).

Generative fMRI synthesis. Closest to us, a small line of work models the stimulus-to-response mapping as a conditional distribution. MindSimulator [22] learns a diffusion prior over fMRI-autoencoder latents and SynBrain [2] a one-step semantic-to-neural VAE mapper; both operate in a compressed, CLIP-conditioned latent space, sharing the two limitations we target (no direct per-ROI readout, a semantic-only bias) and a stability trade-off: the diffusion prior requires multi-sample averaging for stable results, while the one-step mapper is stable but deterministic. StimFlow instead synthesizes in native voxel space with a complementary DINOv2 stream and sidesteps the trade-off via flow matching; we compare against both in Sec. 4.

Visual representations and the semantic bias of CLIP. Which feature space best matches visual cortex remains open. CLIP’s language supervision makes it predictive of high-level, category-selective regions (FFA [14], PPA [6]) but discards low-level structure driving early areas V1–V3, whereas self-supervised transformers such as DINOv2 [5, 26] retain fine spatial structure; supervised and self-supervised models are known to correlate differently with cortical hierarchies [3, 11, 16]. This motivates the multi-stream conditioning we analyse per-ROI in Sec. 4.3, where the two representations prove functionally complementary rather than interchangeable.

Flow matching and diffusion transformers. Flow matching [17] regresses a velocity field onto a probability path; a conditional linear path with independent noise–data coupling yields straight conditional trajectories whose lower curvature allows stable integration with fewer ODE steps than typical diffusion samplers [18, 32]. The Diffusion Transformer (DiT) replaces the U-Net with a trans-

former conditioned via AdaLN [7, 27]. Within neuroimaging, flow and rectified-flow models have so far been applied to medical image synthesis [36] and to brain-connectivity matrices on Riemannian manifolds [4], but to our knowledge not yet to stimulus-conditioned, voxel-space fMRI encoding. StimFlow adapts DiT to 1D fMRI sequences under conditional flow matching to model visual-to-neural mapping in native space, and is among the first to bring CFM to this setting.

3 Method

Two properties of fMRI data motivate our design. First, fMRI voxels are indexed by anatomical location rather than a regular 2D pixel grid, so flattening them into a 1D token sequence imposes far less artificial structure, making a 1D Transformer the appropriate backbone. Second, we want the training and inference distributions to match exactly: conditional flow matching achieves this because the interpolation starts at the same Gaussian distribution inference begins from, and the same ODE is used at both stages. In practice this *train–inference alignment* yields lower-curvature trajectories than a VP-SDE diffusion path [18], allowing stable integration with fewer steps.

3.1 Problem Formulation

We consider the Natural Scenes Dataset (NSD) [1], which provides paired data $\{(\mathbf{x}_i, \mathbf{v}_i)\}_{i=1}^N$ where $\mathbf{x}_i$ denotes a natural image and $\mathbf{v}_i \in \mathbb{R}^{n_s}$ is the corresponding z-scored fMRI response vector for subject s , comprising all voxels in the `nsdgeneral` ROI (n_s ranges from roughly 12 000 to 16 000 across subjects). Because identical stimuli elicit variable responses across trials, this mapping is inherently stochastic; our goal is to learn the conditional distribution $p(\mathbf{v} \mid \mathbf{x})$ directly in *native voxel space*, synthesizing high-fidelity fMRI responses for any image.

3.2 Native-Space Conditional Flow Matching

We cast fMRI synthesis as a conditional flow matching (CFM) problem [17, 18]. Let $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote a pure noise sample in $\mathbb{R}^{n_s}$ and $\mathbf{v}$ the measured fMRI target. We adopt the *conditional linear* probability path (independent coupling [17])

$$\mathbf{x}_t = (1 - t) \mathbf{x}_0 + t \mathbf{v}, \quad t \in [0, 1], \quad (1)$$

whose conditional velocity field is simply $\mathbf{u}_t = \mathbf{v} - \mathbf{x}_0$. A neural network f_θ is trained to predict this velocity, conditioned on the interpolated state $\mathbf{x}_t$, the time t , and a stimulus representation $c(\mathbf{x})$ (described in Sec. 3.3):

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t, \mathbf{x}_0, (\mathbf{x}, \mathbf{v})} \frac{1}{|\Omega|} \sum_{j \in \Omega} w_j \left(f_\theta(\mathbf{x}_t, t, c(\mathbf{x}))_j - (\mathbf{v} - \mathbf{x}_0)_j \right)^2, \quad (2)$$

where Ω is the set of real (non-padded) voxels and w_j is a per-voxel weight set to voxel j 's noise-ceiling reliability, normalised to unit mean over Ω , so that the loss emphasises reliably measured voxels and ignores padding. The reliability is estimated *exclusively from training-set* repeated-stimulus trials (split-half correlation [1]); no test-set data enter the weight computation, so there is no information leakage from evaluation stimuli. The training time t is drawn from a logit-normal schedule [7] that concentrates samples away from the $t \rightarrow 0, 1$ endpoints; its exact parameters are given in Sec. S3.

The linear path choice is deliberate. The conditional interpolant in Eq. (1) is straight by construction; while averaging over the conditional posterior bends the learned *marginal* trajectory, a linear coupling induces lower curvature than a VP-SDE diffusion path [18], letting inference integrate the ODE stably with a modest number of Euler steps ($N_{\text{step}}=50$, an order of magnitude below typical score-based samplers). The whole process runs in native voxel space, so no autoencoder is required and per-ROI analysis can be performed directly on the output.

3.3 Stimulus Conditioning

The visual cortex spans a hierarchy from oriented edges in V1 to semantic categories in areas such as the fusiform face area (FFA). A single CLIP embedding, as used by prior generative encoders, is biased toward the semantic end and underserves early areas, so StimFlow conditions on two complementary streams.

Global semantic context (CLIP). A pooled 1280-d embedding from a frozen SDXL-CLIP encoder, projected by a two-layer MLP and summed with the timestep embedding, forms the global signal $\mathbf{c} = \mathbf{e}_t + \mathbf{e}_{\text{CLIP}}$ that modulates every transformer block through Adaptive Layer Normalization (AdaLN). Because CLIP provides a single scene-level summary, AdaLN, which uniformly scales and shifts each layer's activations, is a natural fit, setting the overall semantic context without attending to specific voxel positions.

Structural context (DINOv2). We extract multi-layer grid features from a frozen DINOv2-ViT-G/14 backbone [5, 26]. Unlike CLIP, these self-supervised features retain spatial structure and fine-grained properties (texture, contour, part-whole relations), spanning low-level structure to emerging semantics across layers. After linear projection and RMSNorm, the token sequence enters each block via multi-head cross-attention, so each voxel patch attends to the most relevant spatial features. Cross-attention is preferred over concatenation or global pooling because the feature map is spatially structured and different cortical regions need different subsets of it, which a single global projection would collapse.

3.4 DiT-1D Velocity Network

The velocity field f_θ is parameterised by a *1D Diffusion Transformer* (DiT-1D), adapted from the DiT architecture [27] to operate on sequential rather than spatial data. The design proceeds in three stages.

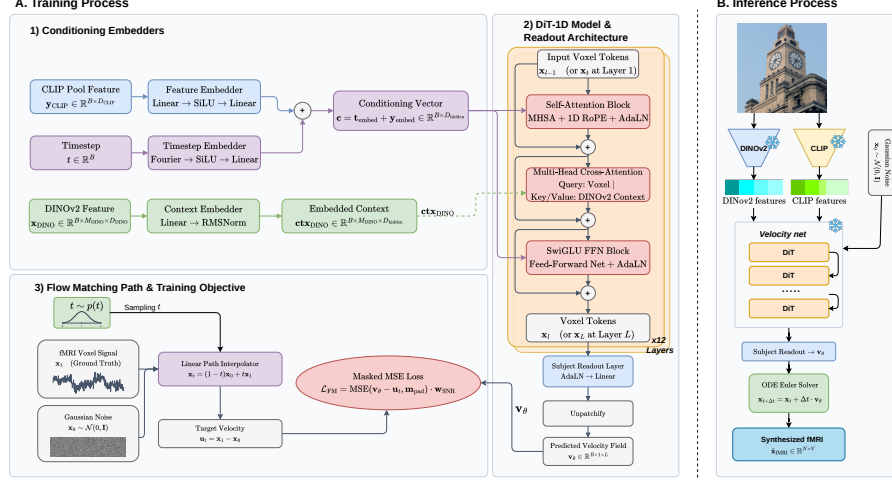


Fig.1: Overview of the StimFlow training and inference framework.

(a) *Training Process*: Conditioning vectors $\mathbf{c}$ (combining CLIP semantics and timestep information) are derived from the input image, while structural context from DINOv2 is cross-attended. Voxel tokens are patchified and processed through 12 DiT-1D layers to regress the conditional velocity field $\mathbf{u}_t$ directly in native voxel space using a masked MSE loss. (b) *Inference Process*: Starting from Gaussian noise $\mathbf{x}_0$, the velocity field is integrated over time using an ODE Euler solver to synthesize the final fMRI response.

Patchification. The z-scored voxel vector $\mathbf{v} \in \mathbb{R}^L$ ($L = 16384$ after zero-padding) is treated as a single-channel 1D signal and tokenised into non-overlapping patches via a strided Conv1d layer, after which fixed sinusoidal positional embeddings encode absolute patch position. The patch width is a low-impact choice fixed to its default; the supporting sweep is in Sec. S2.

Transformer blocks. The trunk stacks $L_d = 12$ transformer blocks. Each block applies QK-normalised multi-head self-attention with Rotary Position Embeddings (RoPE) [33] on queries/keys; modulates the tokens through AdaLN-Zero [27] (per-block scale/shift/gate derived from $\mathbf{c}$, gates zero-initialised so each block starts as identity); injects stimulus structure via a cross-attention layer whose queries are the patch tokens and keys/values the projected DINOv2 tokens; and ends with a SwiGLU [31] feed-forward network, using RMSNorm [37] throughout. These components (RoPE, QK-norm, SwiGLU, RMSNorm, AdaLN-Zero) are standard diffusion-transformer ingredients, not contributions of this work; our focus is their use within native-space flow matching for fMRI.

Unpatchification. A final AdaLN-modulated linear layer maps each patch token back to its corresponding voxel values, and the patches are concatenated to recover the full-length velocity prediction $\hat{\mathbf{u}} \in \mathbb{R}^L$.

3.5 Multi-Subject Shared Trunk with Per-Subject Adapters

To leverage multiple subjects in one model, StimFlow uses a *shared-private* design: all transformer blocks, timestep/CLIP embedders, and context projectors are shared, forming a common trunk that learns subject-invariant stimulus-to-velocity dynamics. The only per-subject parts are two lightweight adapters at the trunk boundaries: a patch embedder (Conv1d) at the input and a readout head (AdaLN + linear) at the output. Since all subjects are zero-padded to a uniform $L = 2^{14} = 16384$, the patch grid is identical regardless of voxel count; mini-batches are subject-homogeneous (each forward pass uses a single adapter pair) and a scalar subject index selects the adapters. This keeps the per-subject parameter count tiny while the shared trunk benefits from all subjects’ data, and, unlike methods that map different subjects’ fMRI into a common latent bottleneck, keeps synthesis in each subject’s native voxel space.

3.6 ROI-Ordered Native Voxel Representation

Before patchification, voxels are sorted once by functional ROI rather than arbitrary index (e.g. V1–V3, hV4, EBA, FFA, PPA), a static reordering that uses no ROI labels at inference. Because generation stays in native voxel space, per-ROI fidelity can be read directly off the synthesized signal by indexing the relevant voxels, without any region-specific decoder or re-mapping step (Secs. 4.3 and 4.4).

3.7 Training and Inference

Training. We optimise the flow-matching loss (Eq. (2)) with AdamW [21] ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.01) using a cosine learning-rate schedule [20] from 10^{-4} to 10^{-5} over 50 epochs, with 500 warm-up steps (see Fig. 1a). Gradients are clipped to unit norm, and all training is conducted in bf16 mixed precision. Effective batch size is $32 \times 8 = 256$ (via gradient accumulation).

Inference. At test time, we solve the learned ODE from $t = 0$ (noise) to $t = 1$ (data) with the Euler method using $N_{\text{step}} = 50$ steps (see Fig. 1b); each noise sample $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ yields a different plausible response. We report a *single-sample* setting ($T=1$) as our primary operating point and an *optional* multi-sample average ($T=N$) approximating the conditional mean. Unlike diffusion-prior encoders that rely on multi-trial averaging and post-selection, StimFlow already attains competitive results at $T=1$ (Tab. 1); averaging is a deliberate variance-reduction option, applied identically to every method for fairness.

4 Experiments

4.1 Experimental Setup

Dataset. We use the Natural Scenes Dataset (NSD) [1], focusing on the four subjects (1, 2, 5, 7) who completed all scanning sessions. Following the standard protocol, each subject has roughly 9,000 unique training images and a shared set of 1,000 test images presented up to three times. We operate on z-scored single-trial betas restricted to the `nsdgeneral` ROI.

Baselines. We compare against two deterministic regression encoders (per-voxel ridge regression and a transformer encoder trained in our pipeline) and two recent generative fMRI synthesizers, MindSimulator [22] and SynBrain [2]. As MindSimulator has no public implementation, we report its published numbers; SynBrain and both regression baselines are run in our pipeline. Every method we run uses the same z-scored single-trial betas, the normalization standard in recent NSD work [22,30], and the identical per-voxel, across-image Pearson metric on the same test split. Re-running SynBrain under this standardized z-scoring yields lower voxel-level Pearson than its original paper, whose preprocessing retains the per-voxel mean.

Metrics. We evaluate at two levels. *Voxel-level* metrics (MSE and Pearson correlation) are computed per voxel across the test images (the standard encoding metric) and averaged over trials; the Pearson correlation is used consistently in all tables and per-ROI analyses. *Semantic-level* metrics are obtained by decoding the synthesized fMRI with a frozen MindEye2 [30] model and comparing the decoded image to the stimulus (Inception, CLIP, EfficientNet, SwAV).

Implementation. StimFlow and its ablation variants (backbone replacements, single-stream variants) are trained with the flow-matching objective and schedule of Sec. 3.7. SynBrain is reproduced from the authors’ released code with its original VAE + S2N objective and hyperparameters; MindSimulator numbers are from the published paper.

4.2 Main Results: Visual-to-fMRI Synthesis

Table 1 shows StimFlow improving over every baseline on all six metrics at both $T=1$ and $T=5$. To isolate the contribution of the flow-matching objective from the rest of our design, we add a *DiT-1D (MSE)* control: the identical backbone and DINOv2+CLIP conditioning, trained with a deterministic per-voxel MSE regression loss instead of conditional flow matching. This control already reaches per-voxel Pearson 0.355 (MSE 0.352), above both generative baselines, which indicates that most of our gain comes from the native-space formulation and the dual conditioning rather than from the generative objective per se. Replacing the MSE loss with conditional flow matching (default StimFlow, $T=1$) adds a further +0.016 in voxel Pearson (0.355→0.371) and a comparable improvement on the semantic metrics; the more decisive benefit of flow matching is its train–inference alignment, which lets a single sample ($T=1$) match what the generative baselines require five draws to achieve (see below). At the voxel level, a single sample reaches Pearson 0.371, a consistent advantage over MindSimulator, with MSE 0.342. Per-subject bootstrap confidence intervals ($B=10,000$) are tight (width ≈ 0.007 ; e.g. subject 1: 0.375 [0.372, 0.378]), so each within-subject estimate is precise; the cross-subject standard deviation (0.051) reflects individual variability (subject 7 has the lowest noise ceiling), not estimation noise. Because $n = 4$ limits formal statistical testing, we report this as a consistent directional advantage across subjects. More saliently, StimFlow at $T=1$ already matches or

Table 1: Visual-to-fMRI synthesis on NSD. StimFlow is the multi-subject model (shared trunk with per-subject adapters, trained jointly on subjects 1, 2, 5, 7); results are averaged over subjects (per-subject Pearson std ≈ 0.05 ; per-subject values in Sec. S6). Voxel-level metrics are computed per voxel across images and averaged over trials; semantic-level metrics are obtained by decoding the synthesized fMRI with a frozen MindEye2 model. Methods are grouped into deterministic regression baselines, single-sample generative ($T=1$), and multi-sample generative ($T=5$). *DiT-1D (MSE)* is an ablative control sharing StimFlow’s backbone and DINOv2+CLIP conditioning but trained with a deterministic MSE loss instead of flow matching, isolating the contribution of the generative objective.

Method	Voxel-Level		Semantic-Level (via decoding)			
	MSE↓	Pearson↑	Incep↑	CLIP↑	Eff↓	SwAV↓
<i>Regression baselines</i>						
Linear Regression	0.394	0.334	90.1%	87.2%	0.728	0.432
Transformer Encoder	0.387	0.337	89.8%	85.5%	0.759	0.440
DiT-1D (MSE)	0.352	0.355	91.0%	89.0%	0.720	0.412
<i>Generative, single sample ($T=1$)</i>						
MindSimulator	0.403	0.346	92.1%	90.4%	0.701	0.396
SynBrain	0.398	0.340	90.8%	88.5%	0.715	0.418
StimFlow	0.342	0.371	94.0%	91.8%	0.668	0.375
<i>Generative, multi-sample ($T=5$)</i>						
MindSimulator	0.385	0.357	93.1%	91.2%	0.689	0.391
SynBrain	0.388	0.352	91.5%	89.4%	0.705	0.410
StimFlow	0.335	0.387	95.0%	92.8%	0.650	0.362
GT fMRI (single-trial)	–	–	96.9%	93.9%	0.604	0.342

exceeds every baseline at $T=5$ on both voxel metrics, indicating the formulation reduces the need for multi-sample averaging. At the semantic level, decoding StimFlow responses with the frozen MindEye2 approaches the GT upper bound (Inception 94.0% vs. 96.9%; CLIP 91.8% vs. 93.9% at $T=1$), and averaging lifts Inception to 95.0%, within two points of GT.

4.3 Backbone Ablation and Feature Complementarity

Conditioning on a structural (DINOv2) and a semantic (CLIP) stream is a deliberate, empirically grounded choice. We justify it with two experiments: a feature-backbone ablation asking *which* features to condition on (Tab. 2), and a per-ROI analysis asking *why* the two streams work better together (Fig. 2). The ablation replaces both conditioning streams with standard supervised backbones (ResNet-50 [12], ConvNeXt [19], ViT-B/16 [5]) and also retains each stream in isolation, at both $T=1$ and $T=5$.

The pairing is deliberate (Tab. 2). Three observations show the choice is not arbitrary. First, the conditioning backbone matters substantially even with the

Table 2: Feature backbone ablation, averaged over subjects 1, 2, 5, 7. The default StimFlow conditions on DINOv2 (structural, via cross-attention) and CLIP (semantic, via AdaLN). Each alternative backbone replaces *both* feature streams. “DINOv2 only” and “CLIP only” retain a single original stream.

Feature backbone	$T=1$		$T=5$	
	MSE↓	Pearson↑	MSE↓	Pearson↑
ResNet-50	0.354	0.348	0.346	0.364
ConvNeXt	0.355	0.345	0.347	0.361
ViT-B/16	0.347	0.363	0.339	0.379
DINOv2 only	0.354	0.351	0.346	0.368
CLIP only	0.361	0.338	0.354	0.351
DINOv2 + CLIP (default)	0.342	0.371	0.335	0.387

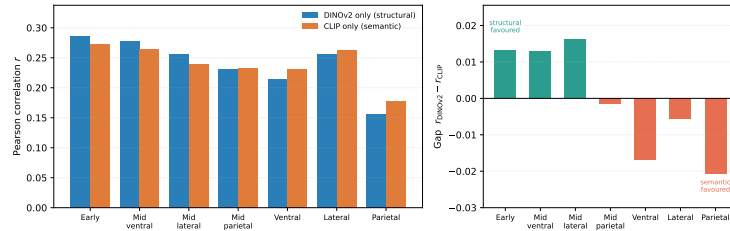


Fig. 2: Per-stream feature complementarity, averaged over subjects 1, 2, 5, 7. (a) Per-voxel Pearson correlation of the DINOv2-only (structural) and CLIP-only (semantic) conditioning variants, by NSD functional stream (early $\rightarrow$ higher). (b) Their gap ($r_{\text{DINOv2}} - r_{\text{CLIP}}$): on average the structural stream leads in early-to-mid visual cortex and the advantage reverses in the higher streams (Spearman $\rho = -0.86$, $p < 0.05$ across the seven streams), suggesting a structure-to-semantics gradient.

synthesis network fixed: $T=1$ per-voxel Pearson ranges from 0.345 (ConvNeXt) to 0.371 (default), a spread of 0.026 comparable to the gaps among methods in Tab. 1. Notably, even a generic ViT-B/16 backbone (0.363) already surpasses both generative baselines (MindSimulator 0.346, SynBrain 0.340), reinforcing the conclusion from Sec. 4.2 that the gain stems primarily from native-space synthesis and dual conditioning rather than the conditioning backbone alone; consistent with the DiT-1D (MSE) control, the flow-matching objective adds a further, more modest voxel-level improvement on top. Second, each original stream in isolation (DINOv2-only 0.351, CLIP-only 0.338) falls below the ViT-B/16 single-backbone result; the dual default recovers the lead (0.371), but the aggregate margin over ViT-B/16 is modest (+0.008) and should not be over-interpreted given four subjects. The stronger evidence for dual-stream conditioning is *mechanistic* rather than aggregate: as the per-ROI analysis below shows, each stream dominates in a different part of the cortical hierarchy, and prior generative encoders that condition *exclusively* on CLIP sacrifice coverage of early visual cortex. That ViT-B/16 alone outperforms both specialist single streams suggests the key requirement is conditioning that spans the full hier-

archy; DINOv2+CLIP is one effective and interpretable instantiation of this principle.

The per-ROI breakdown suggests complementarity (Fig. 2). To understand *why* the pairing helps, Fig. 2 compares the two single-stream variants per NSD functional stream ($T=1$ per-voxel Pearson, averaged over subjects). The DINOv2-only advantage is positive in early-to-mid streams (+0.013 to +0.016 on average) and turns negative in the higher ventral, lateral, and parietal streams (down to -0.021), suggesting a structure-to-semantics gradient along the hierarchy. Across the seven functional streams, the Spearman rank correlation between hierarchical position and the DINOv2–CLIP gap is $\rho = -0.86$ ($p < 0.05$), and all four subjects individually show a negative trend (ρ ranging from -0.21 to -0.93). The per-ROI differences are modest and limited to four subjects, so the crossover should be read as a *descriptive trend consistent with known functional specialisation* rather than a definitive per-region claim; nonetheless, structural features being more informative in early cortex and semantic features in higher areas offers a plausible account of why dual conditioning helps.

4.4 Per-ROI Synthesis Fidelity of the Multi-Subject Model

We next verify that the deployed multi-subject model (shared trunk with per-subject adapters, jointly trained on subjects 1, 2, 5, 7) preserves region-specific structure. Figure 3 reports its per-stream fidelity as the per-voxel Pearson correlation relative to the per-stream noise ceiling. Accuracy remains stable across the hierarchy and recovers a substantial fraction of the ceiling in every stream, with no pronounced weak region, read directly off the generated signal, since synthesis stays in native voxel space.

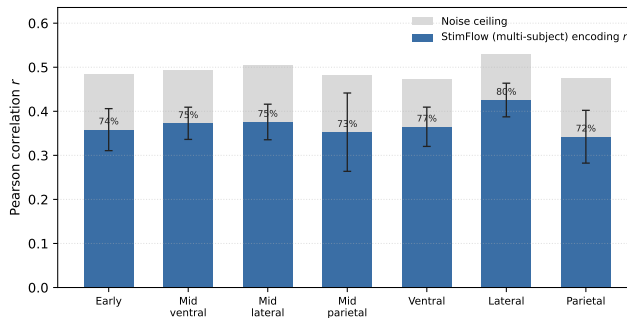


Fig. 3: Per-stream fidelity of the multi-subject model. Per-voxel Pearson correlation of multi-subject StimFlow (shared trunk + per-subject adapters), mean $\pm$ std over subjects 1/2/5/7, overlaid on the per-stream noise ceiling (grey); percentages give the captured fraction of the ceiling. Fidelity remains a substantial fraction of the ceiling across the visual hierarchy, with no region showing a pronounced drop.

Table 3: Few-shot novel-subject adaptation with increasing data budget under a *leave-one-subject-out* protocol. The shared trunk is trained on three subjects and adapted to the held-out subject (only its per-subject adapters + LoRA) using 1–4 h of that subject’s stimulus data. We report per-voxel Pearson correlation (per voxel, across images) for both single-sample ($T=1$) and multi-sample ($T=5$) synthesis. The *Full* column is the multi-subject StimFlow model trained on all four subjects with full data (~ 40 h each); parenthesised values give the fraction of that full-data reference recovered.

Held-out	Train		Full	Few-shot budget			
subject	set	Trials	(ref.)	1 h	2 h	3 h	4 h
Sub1	2,5,7	$T=1$	0.375 (100%)	0.182 (48%)	0.189 (50%)	0.206 (55%)	0.210 (56%)
		$T=5$	0.391 (100%)	0.215 (55%)	0.225 (58%)	0.249 (64%)	0.251 (64%)
Sub2	1,5,7	$T=1$	0.384 (100%)	0.193 (50%)	0.208 (54%)	0.227 (59%)	0.228 (59%)
		$T=5$	0.398 (100%)	0.234 (59%)	0.255 (64%)	0.268 (67%)	0.271 (68%)
Sub5	1,2,7	$T=1$	0.423 (100%)	0.260 (61%)	0.271 (64%)	0.281 (66%)	0.290 (68%)
		$T=5$	0.437 (100%)	0.294 (67%)	0.314 (72%)	0.326 (74%)	0.331 (76%)
Sub7	1,2,5	$T=1$	0.301 (100%)	0.166 (55%)	0.170 (56%)	0.176 (58%)	0.175 (58%)
		$T=5$	0.320 (100%)	0.193 (60%)	0.206 (64%)	0.212 (66%)	0.213 (67%)
Average	–	$T=1$	0.371 (100%)	0.200 (54%)	0.210 (57%)	0.222 (60%)	0.226 (61%)
		$T=5$	0.387 (100%)	0.234 (60%)	0.250 (65%)	0.264 (68%)	0.267 (69%)

4.5 Cross-Subject Generalization and Few-Shot Adaptation

Finally, we evaluate cross-subject generalization under a leave-one-subject-out protocol. The shared trunk is trained on three subjects and adapted to the held-out one by fitting only its per-subject input/output adapters and a low-rank (LoRA) residual [13] on the last block (~ 0.62 M trainable parameters, under 1% of the trunk). Scaling the held-out subject’s data from 1 to 4 hours (one hour ≈ 750 trials), Tab. 3 reports the per-voxel Pearson correlation and the fraction of the full-data reference recovered, for $T=1$ and $T=5$; the reference is the multi-subject model trained on all four subjects with full ~ 40 h each (Tabs. 1 and 2). At $T=5$, adaptation recovers $\sim 60\%$ from one hour and $\sim 69\%$ at 4 h; even $T=1$ recovers $\sim 54\%$ at 1 h, showing the shared trunk transfers to unseen subjects with limited data.

The recovered fraction (Tab. 3) rises steadily with budget for every held-out subject, steepest in the first hour, still gaining through 4 h rather than saturating, and multi-sample averaging adds a further 6–8 points. That one hour recovers over half the full-data reference by adapting only two per-subject linear layers and a low-rank residual indicates the shared trunk captures much of the transferable visual-to-neural structure.

4.6 Qualitative Analysis

Reconstruction comparison. Figure 4 shows the semantic-level evaluation qualitatively: reconstructions decoded (via frozen MindEye2) from StimFlow-synthesized

versus real fMRI. Across diverse content (a surfer, skiers, a giraffe, a steam locomotive), the synthesized reconstructions recover object identity, pose, layout, and dominant colour. As expected, real-fMRI reconstructions remain strongest and bound what any synthesized signal can achieve under the same decoder. Among synthesized variants, $T=5$ is clearly better (cleaner, sharper), while $T=1$ recovers category and coarse layout but is noisier, already capturing the dominant semantics that averaging sharpens rather than introduces. The hardest case is the indoor bathroom scene, whose fine spatial layout is only partially recovered, in line with the general difficulty of fine-grained spatial structure relative to dominant-category content.



Fig. 4: Qualitative reconstruction comparison (subject 1). Each column is one test image. Rows, top to bottom: the seen stimulus (GT), the MindEye2 reconstruction from ground-truth fMRI, and the reconstructions from StimFlow-synthesized fMRI at $T=1$ and $T=5$. Reconstructions from synthesized fMRI recover the same semantic content as those from real fMRI.

Concept-selective region localization. A key promise of generative encoding is localizing concept-selective cortex from synthesized responses without new scans [22]. Grouping the shared test images by caption concept, we average the synthesized responses per group and subtract the all-image mean to form a per-voxel selectivity map on the subject-1 flatmap (Fig. 5). The four concepts produce *distinct*, anatomically sensible signatures: *person* in body/face-responsive cortex (EBA, FFA), *vehicle* in scene/place regions (PPA, OPA), and *food/animal* in ventral-temporal cortex near FFA/hV4. The synthesized maps place their peaks in the same ROIs as the ground truth, remaining sparse and focal with only minor differences in extent. Quantitatively, the Pearson correlation between synthesized and ground-truth selectivity maps (per concept across all voxels, averaged over concepts and subjects) is $r = 0.94$ (range 0.88–0.98) with a top-10% Dice overlap of 0.81, confirming the agreement holds quantitatively, not only by impression.

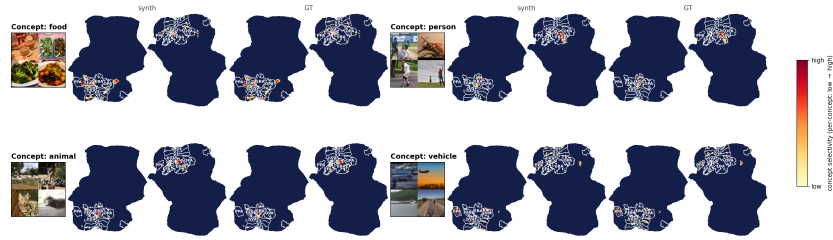


Fig. 5: Concept-selective region localization (subject 1). For each concept (rows), a montage of example stimuli is shown beside the selectivity map (mean response over the concept’s images minus the average response across all test images) rendered on the cortical flatmap, from StimFlow’s synthesized responses (*synth*) and the ground-truth responses (*GT*). White contours mark functional ROIs (V1–hV4, FFA/OFA, PPA/OPA, EBA, VWFA). Each concept is scaled to its own range. The synthesized maps localize the same concept-selective regions as the ground truth.

5 Conclusion

Three findings emerge from our experiments. First, the primary practical benefit of the generative formulation is *train–inference alignment*: because the flow-matching path begins at the same Gaussian distribution used at test time, a single sample already matches or exceeds multi-sample baselines, reducing the need for post-hoc averaging or selection. The model does *not* yet capture meaningful trial-to-trial variability: averaging further improves every metric, indicating that the inter-sample variance is estimation noise rather than neural signal. Single-sample fidelity remains high precisely because training and inference share the same starting distribution. Second, generating in native voxel space, rather than in a compressed latent code, unlocks analyses that prior generative encoders cannot perform directly: per-ROI fidelity readout, concept-selective region localization, and cross-subject adaptation via lightweight adapters, all without an auxiliary decoder. Third, the structural (DINOv2) and semantic (CLIP) conditioning streams show a consistent complementary trend along the cortical hierarchy ($\rho = -0.86$, $p < 0.05$), and their pairing yields gains over either stream alone while matching or slightly exceeding generic backbone replacements; the aggregate margin is modest, but the per-ROI mechanistic rationale provides a principled basis for the dual-stream design.

Limitations and future work. StimFlow approximates the conditional mean of $p(\mathbf{v} \mid \mathbf{x})$; it does not yet model the structured component of trial-to-trial variability (e.g. cognitive-state or attention-dependent fluctuations), which remains an open direction. Performance is bounded by frozen backbones ($\sim 72\text{--}80\%$ noise ceiling), and validation is limited to four NSD subjects; broader evaluation across datasets, tasks, and populations is needed before generated responses could supplement, let alone substitute for, empirical data in applied or clinical settings.

References

1. Allen, E.J., St-Yves, G., Wu, Y., Breedlove, J.L., Prince, J.S., Dowdle, L.T., Nau, M., Caron, B., Pestilli, F., Charest, I., et al.: A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience* **25**, 116–126 (2022)
2. Chen, C., Wu, Z., Zhang, Y., Yao, J., Zou, J., Xu, R., Lin, Z., Du, C., He, H.: SynBrain: Synthetic brain signal generation for learning neural representations. arXiv preprint arXiv:2405.10690 (2024)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
4. Collas, A., Ju, C., Salvy, N., Thirion, B.: Riemannian flow matching for brain connectivity matrices via pullback geometry. In: NeurIPS (2025)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
6. Epstein, R., Kanwisher, N.: A cortical representation of the local visual environment. *Nature* **392**, 598–601 (1998)
7. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorber, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: ICML (2024)
8. Gifford, A., Lahner, B., Saba-Sadiya, S., Vilas, M.G., Lascelles, A., Oliva, A., Kay, K., Roig, G., Cichy, R.M.: The Algonauts project 2023 challenge: How the human brain makes sense of natural scenes. arXiv preprint arXiv:2301.03198 (2023)
9. Güçlü, U., van Gerven, M.A.J.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience* **35**(27), 10005–10014 (2015)
10. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**(5539), 2425–2430 (2001). <https://doi.org/10.1126/science.1063736>
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 16000–16009 (2022)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
14. Kanwisher, N., McDermott, J., Chun, M.M.: The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**(11), 4302–4311 (1997)
15. Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L.: Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008)
16. Khaligh-Razavi, S.M., Kriegeskorte, N.: Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology* **10**(11), e1003915 (2014). <https://doi.org/10.1371/journal.pcbi.1003915>
17. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
18. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR (2023)

19. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: CVPR. pp. 11976–11986 (2022)
20. Loshchilov, I., Hutter, F.: SGDR: Stochastic gradient descent with warm restarts. In: ICLR (2017)
21. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
22. Luo, L., Du, C., Li, J., He, H.: MindSimulator: Imitating the human brain to synthesize fMRI under visual stimuli. *ACM Trans. Multimedia Comput. Commun. Appl.* (2024)
23. Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., Just, M.A.: Predicting human brain activity associated with the meanings of nouns. *Science* **320**(5880), 1191–1195 (2008). <https://doi.org/10.1126/science.1152876>
24. Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L.: Encoding and decoding in fMRI. *NeuroImage* **56**(2), 400–410 (2011)
25. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., Gallant, J.L.: Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology* **21**(19), 1641–1646 (2011). <https://doi.org/10.1016/j.cub.2011.08.031>
26. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. *TMLR* (2024)
27. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
29. Scotti, P.S., Banerjee, A., Goode, J., Shabalin, S., Nguyen, A., Cohen, E., Dempsey, A.J., Verlinde, N., Yundler, E., Weiber, D., et al.: Reconstructing the mind’s eye: fMRI-to-image with contrastive learning and diffusion priors. In: NeurIPS (2023)
30. Scotti, P.S., Tripathy, M., Torrico, C.K., Kneeland, R., Chen, T., Narang, A., Santhirasekaran, C., Hillman, J., Khosla, M., Naselaris, T., et al.: MindEye2: Shared-subject models enable fMRI-to-image with 1 hour of data. In: ICML (2024)
31. Shazeer, N.: GLU variants improve transformer. *arXiv preprint arXiv:2002.05202* (2020)
32. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
33. Su, J., Ahmed, M., Lu, Y., Pan, S., Wen, B., Liu, Y.: RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing* **568**, 127063 (2024)
34. Takagi, Y., Nishimoto, S.: High-resolution image reconstruction with latent diffusion models from human brain activity. In: CVPR. pp. 14453–14463 (2023)
35. Yamins, D.L.K., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., DiCarlo, J.J.: Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**(23), 8619–8624 (2014)
36. Yazdani, M., Medghalchi, Y., Ashrafian, P., Hacıhaliloglu, I., Shahriari, D.: Flow matching for medical image synthesis: Bridging the gap between speed and quality. In: *Medical Image Computing and Computer Assisted Intervention (MICCAI)*. LNCS, vol. 15975, pp. 216–226. Springer (2025)
37. Zhang, B., Sennrich, R.: Root mean square layer normalization. In: NeurIPS (2019)

Supplementary Material

This supplementary material reports the design sweeps that justify the hyperparameters of StimFlow. We separate every choice into *high-impact* settings, reported in the main text because they shape fidelity, and *low-impact* settings, fixed to robust defaults and tabulated here so the main architecture description stays focused. Four axes are covered: the DiT-1D architecture (Sec. S2), the training loss and time schedule (Sec. S3), the inference solver budget (Sec. S4), and the few-shot LoRA adapter (Sec. S5).

S1 Protocol

Each sweep varies a *single* hyperparameter while holding all others at their default values (patch width $P=32$, token width $D=512$, attention heads $H=8$, depth $L_d=12$, sequence length $L=16384$). All variants use the default dual-stream conditioning (DINOv2 + CLIP) and the multi-subject shared-trunk model trained jointly on subjects 1, 2, 5, and 7 with the optimisation schedule of the main text (§3.7). We report voxel-level MSE and Pearson correlation (per voxel, across images) at the single-sample operating point ($T=1$), averaged over the four subjects; the default configuration reproduces the main-paper number (Pearson 0.371, MSE 0.342). Parameter counts are for the deployed multi-subject model: the shared trunk (71.9M at the default setting) plus four per-subject adapter pairs (≈ 0.56 M each).

S2 DiT-1D architecture

Low-impact: patch width and attention heads (Tab. S1). Both are minor relative to depth and are therefore omitted from the main text. Head count only re-partitions a fixed token width and is essentially constant (within 0.004 Pearson). Patch width trades token count (hence self-attention cost) against voxel granularity and stays flat near the default ($P=16$ vs. 32 within 0.003), dropping only at the coarsest setting ($P=64$, -0.014); this confirms that the per-ROI readout of the main text does *not* rely on any patch-to-region alignment. Both leave the parameter count unchanged (≈ 74 M), so the defaults are robust rather than tuned.

High-impact: trunk depth (Tab. S2). Depth is the one architecture knob with a first-order effect: $6 \rightarrow 12$ blocks adds 0.019 Pearson, after which accuracy plateaus ($12 \rightarrow 24$ adds ≤ 0.001 at $2\times$ the parameters). We therefore expose $L_d=12$ in the main text (the depth at which the trunk is expressive enough without over-parameterising).

Table S1: Low-impact architecture sweeps ($T=1$, averaged over subjects 1/2/5/7). Each block varies one hyperparameter with the rest at default; the default row (MSE 0.342, Pearson 0.371) is the same model, repeated as the anchor for each axis. N is the token count, d the per-head dimension.

Hyperparameter	Value	MSE↓	Pearson↑
Patch width P	16 ($N=1024$)	0.341	0.368
	32 (default)	0.342	0.371
	64 ($N=256$)	0.349	0.357
Heads H	4 ($d=128$)	0.344	0.367
	8 (default)	0.342	0.371
	16 ($d=32$)	0.343	0.369

Table S2: Trunk-depth sweep ($T=1$, averaged over subjects 1/2/5/7). Default in bold.

Depth L_d	Params (M)	MSE↓	Pearson↑
6	39.4	0.351	0.352
12 (default)	74.1	0.342	0.371
18	108.8	0.341	0.372
24	143.4	0.342	0.371

S3 Loss and time-schedule settings

For completeness we record the exact settings behind the flow-matching objective (Eq. (2)); these are standard choices that we do not tune. The training time t follows a logit-normal schedule, $\text{logit}(t) \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu=-1$ and $\sigma=1$, warped by the resolution-aware shift $t \mapsto S t / (1 + (S-1)t)$ with S tied to the sequence length. The per-voxel weight w_j is the voxel’s noise-ceiling reliability, normalised to unit mean over the real (non-padded) voxels, and the loss is masked to those voxels (both as defined in the main text).

S4 Inference solver budget

At test time we integrate the learned ODE with a fixed-step solver (Fig. S1). Euler fidelity saturates by 50 steps; halving to 25 costs only 0.006 Pearson while doubling to 100 adds nothing. A second-order Heun solver reaches the same plateau in 25 steps, but since it spends two network evaluations per step this is the *same* 50 function evaluations as Euler-50, i.e. no gain at equal compute. We therefore report the simpler Euler, 50-step setting in the main text and treat the solver as a non-sensitive inference knob. This sweep is inference-only: all points reuse the single default checkpoint.

S5 Few-shot LoRA adaptation

The few-shot protocol (Sec. 4.5) adapts the frozen shared trunk to a held-out subject with its per-subject adapters plus a LoRA residual on the last trunk

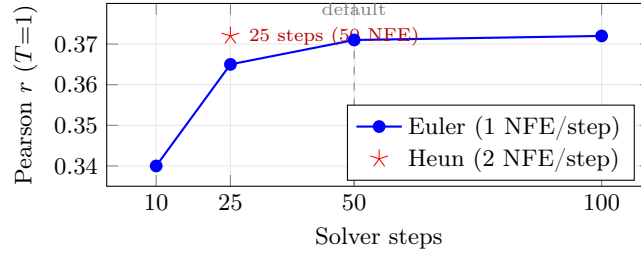


Fig. S1: Inference solver vs. step count ($T=1$, averaged over subjects 1/2/5/7). Euler fidelity (blue) plateaus by 50 steps (default, dashed line). Heun (red star) reaches the same plateau in 25 steps, but requires the same 50 NFEs as Euler-50 (2 NFEs/step), with no gain at equal compute.

block. Table S3 sweeps the LoRA rank and the number of adapted trailing blocks at the 1 h budget ($T=5$, averaged over the four held-out subjects). Adapters alone (no LoRA) leave the frozen trunk unable to specialise; a rank-8 residual on a single block recovers most of the gain, after which larger rank or more blocks add little for roughly double the trainable parameters. We thus fix rank 8 on the last block, adapting only its attention and feed-forward Linear layers while norms and AdaLN conditioning stay frozen.

Table S3: LoRA rank and adapted-block sweep for few-shot adaptation (1 h budget, $T=5$, averaged over the four held-out subjects). “Trainable” counts the per-subject adapters (0.56 M) plus the LoRA residual. Default in bold.

Adapted blocks	Rank	Trainable (M)	Pearson $\uparrow$
none (adapters only)	—	0.559	0.204
last 1	4	0.592	0.226
last 1	8	0.625	0.234
last 1	16	0.690	0.235
last 2	8	0.690	0.236

S6 Per-subject voxel-level statistics

Table S4 reports the per-subject per-voxel Pearson correlation of the default StimFlow multi-subject model (shared trunk + per-subject adapters) together with bootstrap 95% confidence intervals ($B=10,000$). Within each subject the estimate is highly precise (CI width ≈ 0.007), confirming that the voxel-level metric is not driven by estimation noise. The cross-subject standard deviation (0.051) reflects genuine individual differences in data quality and noise ceiling: subject 5 has the highest noise ceiling and the best performance, while subject 7 has the lowest.

Multi-sample averaging ($T=5$) significantly improves over single-sample ($T=1$) within every subject (paired bootstrap $p < 10^{-4}$, $\Delta \approx +0.016$ per subject, $B=10,000$).

Table S4: Per-subject voxel-level Pearson correlation (mean over voxels, computed across images) with bootstrap 95% CI ($B=10,000$). Default multi-subject StimFlow model.

Subject	N_{voxels}	$T=1$ Pearson [95% CI]	$T=5$ Pearson [95% CI]
1	15 724	0.375 [0.372, 0.378]	0.391 [0.388, 0.394]
2	14 278	0.384 [0.380, 0.388]	0.398 [0.394, 0.402]
5	13 039	0.423 [0.419, 0.427]	0.437 [0.434, 0.441]
7	12 682	0.301 [0.297, 0.304]	0.320 [0.317, 0.323]
Average	–	0.371 $\pm$ 0.051	0.387 $\pm$ 0.049

S7 The SynBrain Baseline: Why Its Reported Per-Image Pearson Is Inflated

This section explains why SynBrain’s originally reported Pearson (0.687) is far above the comparable value (0.340). The effect is specific to the *per-image, across-voxel* Pearson (the metric SynBrain reports) combined with its normalization, and is therefore analyzed here in that metric; it is distinct from the per-voxel, across-image encoding Pearson used for all results in the main text.

Two normalization regimes. The main text z-scores the single-trial betas: each voxel is centred and scaled to zero mean and unit variance using training-set statistics, following the standard NSD encoding/decoding protocol [1, 22, 30]. The publicly released SynBrain pipeline instead divides every beta by a fixed constant (2000) applied uniformly to all voxels, *without* removing the per-voxel mean. We denote the two regimes *mean-retained* and *z-scored*.

SynBrain’s metric, and why the normalization matters. SynBrain scores fidelity with the per-image, across-voxel Pearson correlation: for test image i it correlates the synthesized voxel pattern $\hat{\mathbf{v}}_i \in \mathbb{R}^V$ against the measured pattern $\mathbf{v}_i$ across the V voxels, then averages over images. Crucially, this correlation is invariant to any positive rescaling, $\text{PCC}(a\hat{\mathbf{v}}_i, \mathbf{v}_i) = \text{PCC}(\hat{\mathbf{v}}_i, \mathbf{v}_i)$ for $a > 0$, so the constant division by 2000 *by itself cannot change this metric*. The regimes differ in two ways — z-scoring removes the per-voxel mean and additionally rescales each voxel by its own standard deviation — but, as the mean-only control below isolates, it is the retained *per-voxel mean*, not the rescaling, that drives the inflation: under across-voxel scoring with the mean kept, a model only has to reproduce the static per-voxel mean pattern to attain a high score, so the reported Pearson is largely an artifact of the normalization rather than evidence of stimulus-specific fidelity.

Why mean retention inflates per-image Pearson. Write each measured pattern as $\mathbf{v}_i = \bar{\mathbf{m}} + \mathbf{r}_i$, where $\bar{\mathbf{m}}$ is the static per-voxel mean (shared by every image and reflecting baseline and anatomical offsets) and $\mathbf{r}_i$ is the stimulus-driven residual. Under the mean-retained regime $\bar{\mathbf{m}}$ is present in both the target and any reasonable prediction, and because it is identical across images it is trivially reproduced. The across-voxel correlation is then dominated by agreement on $\bar{\mathbf{m}}$

rather than on the stimulus-specific $\mathbf{r}_i$, inflating per-image Pearson. Z-scoring removes $\bar{\mathbf{m}}$, leaving $\mathbf{r}_i$ alone, so the metric measures stimulus-driven fidelity.

A mean-only control isolates the effect. Consider a predictor that ignores the stimulus and outputs the constant per-voxel training mean $\bar{\mathbf{m}}$ for every image. Under the mean-retained regime this trivial predictor already attains a high per-image Pearson of 0.83 on subject 1, *exceeding* SynBrain’s reported 0.687 (Fig. S2, “mean-only”), despite carrying no stimulus information; under z-scoring its correlation collapses to chance, because the prediction equals the train mean and loses all across-voxel variance once the mean is removed. Since the static mean alone already yields 0.83 in the mean-retained regime — itself above SynBrain’s reported 0.687 — a score reported in that regime reflects mostly this mean rather than stimulus-specific fidelity; removing the mean by z-scoring strips the artifact and gives the comparable 0.340.

MSE moves oppositely. Unlike Pearson, mean-squared error is scale-dependent: the constant division shrinks the signal’s dynamic range, so the mean-retained regime also reports a much smaller MSE (0.139 vs. 0.398 after z-scoring). The two metrics improve together only because both follow from the same normalization, not because the synthesized signal is closer to the target.

Convergent evidence for the comparable value. Under z-scoring, our re-run of SynBrain (0.340) closely matches the independently reported MindSimulator value (0.346); only SynBrain’s mean-retained number (0.687) stands apart. Two independently developed methods converging to ≈ 0.34 under the standard normalization, with the single outlier accounted for by mean retention, supports treating 0.340 as the comparable voxel-level score.

Reproduction protocol. The effect is reproduced as follows, using the released SynBrain checkpoint and the shared 1,000-image test set restricted to the `nsdgeneral` ROI: (i) obtain SynBrain’s synthesized betas $\hat{\mathbf{v}}_i$ and the measured betas $\mathbf{v}_i$; (ii) compute per-image across-voxel Pearson under the mean-retained regime (constant division, no mean removal) and under z-scoring (per-voxel train mean/std); (iii) add the mean-only control by replacing $\hat{\mathbf{v}}_i$ with the broadcast per-voxel train mean; (iv) report the mean over images. The mean-only control in Fig. S2 is computed on the measured betas of subject 1, for which the mean-retained signal is available; the analysis script is released with our code.

Implication for evaluation. Semantic-level metrics alone can mask voxel-level error: a robust frozen decoder can recover the correct image even from a coarsely synthesized response, so high decoded-image scores do not guarantee faithful voxel patterns. A voxel-level Pearson computed under standard z-scoring is therefore a necessary check, and we report both levels throughout.

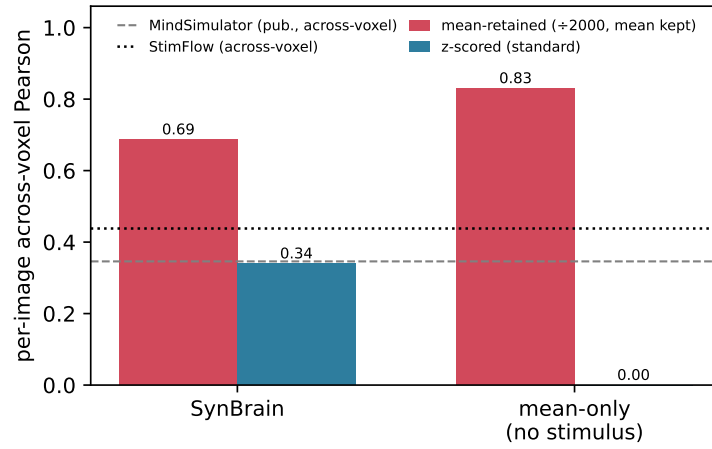


Fig. S2: Per-image, across-voxel Pearson under two normalization regimes. Per-image, across-voxel Pearson (SynBrain’s reported metric) for SynBrain and for a stimulus-agnostic mean-only predictor, under the mean-retained regime (constant division, per-voxel mean kept) and under standard z-scoring. MindSimulator (published) and StimFlow are shown for reference. The mean-only predictor scores high when the mean is retained and collapses under z-scoring, showing that the inflated mean-retained Pearson reflects the static per-voxel mean rather than stimulus-driven fidelity. The mean-only bars are computed on subject 1 (the subject with the mean-retained signal on hand); SynBrain bars are from our reproduction (mean-retained = originally reported 0.687; z-scored = 0.340).