BRAINWM: A MULTIMODAL WORLD MODEL FOR REPRESENTING, FORECASTING, AND SIMULATING WHOLE-BRAIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Most brain imaging foundation models treat entire fMRI scans as static images, yet the brain is a dynamic world: its state is described by both structure and function, and it evolves, driven by intrinsic spontaneous activity, external stimuli, and long-term processes such as aging. We present BrainWM, a multimodal world model for whole-brain 4D dynamics that unifies representation learning, forecasting, and simulation in a single model with three components: BrainWM-VAE encodes structural and functional MRI into an integrated structure–function latent state; BrainWM-DiT learns to generate the next state from the past in this latent space via flow matching; and BrainWM-BAM learns latent actions for state transitions through prior–posterior modeling. BrainWM is trained progressively in three stages: first unified latent-space learning, then within-run generative forecasting pretraining, and finally post-training for conditional cross-run simulation. On large-scale internal and external test sets, BrainWM achieves performance on downstream tasks that surpasses or is comparable to that of foundation models designed specifically for representation learning, while also being capable of forecasting and simulating functional and structural brain changes—tasks that these models are not designed to address. To our knowledge, BrainWM is the first brain imaging world model systematically evaluated for representing, forecasting, and simulating whole-brain dynamics.

1 INTRODUCTION

Functional magnetic resonance imaging (fMRI) is the primary non-invasive tool for observing activities in the human brain. With large-scale neuroimaging datasets and self-supervised pretraining, brain foundation models have developed rapidly and achieved strong performance on various tasks ranging from demographic prediction to clinical diagnosis (Ghamizi et al., 2026). The brain, however, is not static but a dynamic world system (Figure 1). The state of a brain is described by its function, constrained by structural anatomy, and evolves over time, driven by intrinsic spontaneous activity, external tasks and stimuli, and long-term physiological changes like aging. Therefore, beyond representing the current brain state, it is equally valuable to forecast its future state and to simulate its state under different conditions, a view that matters both for understanding brain mechanisms and in the clinic. Consider a psychiatric disorder: brain representation supports discriminative diagnosis, while forecasting anticipates its natural progression and simulation estimates the effect of a treatment. These tasks require a world model specifically designed to capture the temporal and conditional dynamics of brain states.

Brain foundation models have made steady progress, from region-level modeling to whole-brain voxel modeling (Caro et al., 2024; Wang et al., 2026a; 2025; 2026b), and from fMRI (functional) alone to integrating sMRI (structural) (Dong et al., 2025). Yet most of these models treat an fMRI scan as a static image, learning a holistic representation through masked reconstruction. Learning brain dynamics, instead, calls for generative methods that model how a brain state evolves to the next step or under a different condition. Generative models have recently been introduced to brain imaging (Seo et al., 2025; Duan et al., 2026; Xia et al., 2026a), but this line of work has so far focused on improving the quality of synthesized samples or learning transferable representations, rather than modeling dynamic trajectories. Concurrently with this work, several studies have begun to generate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

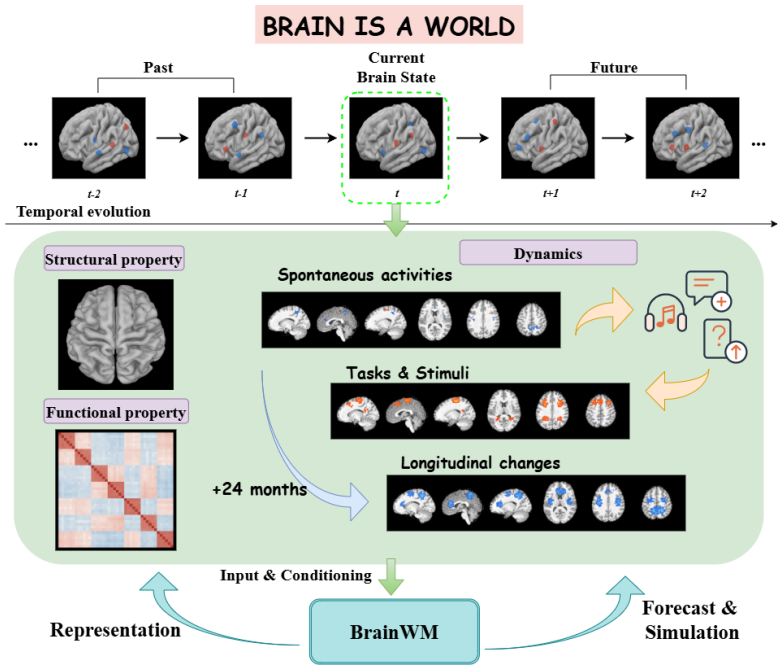

Figure 1: **The brain is a world.** *Top:* brain states evolve over time. *Middle:* a state is described by both structure and function; spontaneous activity, tasks and stimuli, and long-term changes drive its evolution. *Bottom:* BrainWM takes these as inputs and provides representation, forecasting, and simulation from a single model.

brain dynamics (Xia et al., 2026b; Dong et al., 2026). A brain world model, however, demands more than a better generator, and several gaps remain. First, structural information is currently treated primarily as a static condition, rather than as part of the brain state itself. Second, no model yet explicitly models the transition dynamics of brain states beyond next-state prediction. Third, there is no comprehensive training and evaluation framework that develops and assesses brain-state representation, future-state forecasting, and conditional simulation within a single model.

We propose BrainWM (Figure 1, bottom), a multimodal world model for whole-brain 4D dynamics with three components designed to address these gaps. First, BrainWM-VAE (Variational Autoencoder) encodes sMRI and fMRI with modality-specific 3D encoders and fuses them through a shared spatial posterior into a single structure–function latent space, so that a subject's anatomy and function are unified in one state and structure becomes generable rather than a static condition. Second, built atop the integrated latent representation, BrainWM uses a DiT (Diffusion Transformer) as the generative backbone, together with a BAM (Brain Action Model) that learns state transitions in the inverse direction: a posterior summarizes the change that actually occurred between adjacent brain states, while a prior infers the change that should occur given conditions such as stimuli, age interval, and conditions specified in scripts. During training, the prior and posterior are aligned, and at inference, the DiT uses the prior's action tokens as the prompt to generate next state. Finally, BrainWM is trained in three stages: reconstruction-based training for unified latent-space learning, generative pretraining for next-window forecasting, and post-training for cross-condition simulation. Together, these designs allow BrainWM to represent structure and function as a unified brain state, learn how that state changes under different conditions,

and use the learned dynamics for both forecasting and simulation within a single framework.

Validated on large-scale cohorts, BrainWM achieves downstream-task performance comparable to foundation models designed specifically for brain representation, while also exhibiting strong generative capabilities in forecasting and simulation. Our contributions are threefold:

- **A unified multimodal brain state.** BrainWM encodes structural and functional images into a unified structure–function latent state that fuses structural and functional information, making structure generable rather than merely a condition.
- **A generative world model of brain dynamics.** Using the DiT and BAM architecture and three-stage training, BrainWM learns the transition rules of brain states in latent space, enabling forecasting and conditional simulation from a given state.
- **A three-capability world model.** To our knowledge, BrainWM is the first brain world model to span representation, forecasting, and simulation and to be evaluated across all three capabilities within a comprehensive framework.

2 Related Work

Foundation Models of Brain MRI: Early brain foundation models such as BrainLM and Brain-JEPA were pretrained on ROI-level time series, while BrainMass uses functional connectivity matrices (Caro et al., 2024; Dong et al., 2024; Yang et al., 2024). NeuroSTORM and SLIM-Brain pretrain directly on 4D voxel data to preserve spatial detail (Wang et al., 2026a; 2025). In terms of modality, BrainIAC is trained on sMRI (Tak et al., 2026), and Brain Harmony unifies structural and functional information into 1D tokens (Dong et al., 2025). These models learn static brain representations for downstream prediction rather than modeling how brain states evolve. This is beginning to change: Seo et al. (2025) were the first to use a diffusion transformer for conditional generation of voxel-level 4D task fMRI, and FM-fMRI synthesizes task-state ROI series from resting state (Duan et al., 2026), but neither takes rolling brain states forward as its target. Concurrently with this work, BrainWorld and NeuroWorld do roll brain states forward in latent space, yet neither supports cross-condition simulation nor a unified structure–function state (Xia et al., 2026b; Dong et al., 2026).

Latent Action Learning for World Models: Latent action modeling originates in embodied intelligence, where actions describe physical behaviors that change an agent's state. It infers those actions in latent space from consecutive observations through inverse dynamics and uses them to predict the next state through forward dynamics (Schmidt & Jiang, 2024; Bruce et al., 2024). To enable generation without future observations at inference time, Being-H0.7 uses a posterior with access to the future to supervise a prior conditioned only on history and instructions (Luo et al., 2026). We extend latent action learning to brain imaging with BAM. Here, actions are defined as the conditions driving brain-state transitions. BAM infers these actions from observed transitions and learns to predict them from spontaneous activity, external stimuli, and long-term physiological changes, using the predicted actions to guide forecasting and simulation.

3 Methods

BrainWM (Figure 2) compresses whole-brain sMRI and 4D fMRI into a unified VAE latent space and learns latent brain dynamics through flow matching (Lipman et al., 2023), using a DiT backbone (Peebles & Xie, 2023) and a Brain Action Model (BAM). Three-stage training enables unified latent learning and conditional brain generation.

3.1 BrainWM-VAE

BrainWM-VAE (Figure 2a) is a multimodal 3D variational autoencoder (Kingma & Welling, 2014) that takes a subject's paired structural and functional images as input. Two 3D convolutional encoders encode the structural image x^s once and the functional images x_t^f frame by frame, compressing each 96^3 volume to a 12^3 grid. The structural embedding is broadcast over time and concatenated with each frame's functional embedding. A shared head outputs the posterior parameters $\mu_t,\ \log \sigma_t^2 \in \mathbb{R}^{32\times 12^3}$, from which $z_t \sim \mathcal{N}(\mu_t, \sigma_t^2)$ is sampled. Two decoders reconstruct $\hat{x}^s$ and $\hat{x}_t^f$ from the same z_t, so the latent must preserve both the subject's anatomy and function at time t (Appendix A). The VAE is then frozen, and the mean μ_t are extracted as latents for later phases.

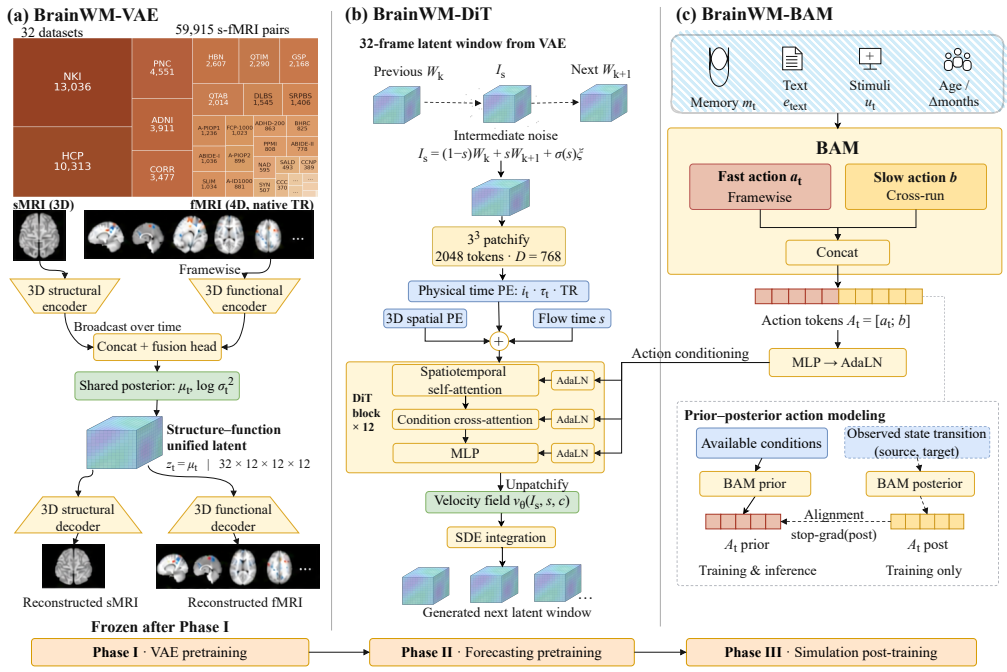

Figure 2: **Overview of BrainWM.** (a) BrainWM-VAE encodes paired sMRI and fMRI into a unified structure–function latent with two 3D encoders and a shared posterior. (b) BrainWM-DiT takes a 32-frame latent window with spatial, physical-time and flow-time embeddings and predicts the velocity field $v_\theta(I_s, s, c)$ transporting it to the next, with $\sigma(s) = \epsilon(1-s)\sqrt{s}$ the interpolant's noise amplitude. (c) BrainWM-BAM compresses memory, text, stimuli and age into fast and slow action tokens that condition the DiT through AdaLN; a prior predicts them from conditions, supervised by a posterior that sees the observed transition. Bottom: the three training phases.

3.2 BrainWM-DiT

BrainWM-DiT (Figure 2b) generates the next 32-frame latent window from the previous window and the conditions. It is a 12-block transformer over patches of size 3^3 taken from each frame, where every block applies spatiotemporal self-attention across the whole window, cross-attention to the text condition tokens, and an MLP. All three sublayers are modulated by per-frame AdaLN, through which BAM action tokens are injected (Section 3.3). Given the interpolant state I_s and flow time s, BrainWM-DiT outputs the flow-matching velocity field $v_\theta(I_s, s, c)$. Since resampling to a common TR would alter the original temporal dynamics, we maintain each cohort's native TR and apply an explicit physical-time embedding alongside the spatial and flow-time embeddings. It encodes each frame's index, the TR, the acquisition time from the start of the run, and Fourier features of that time (Appendix A).

3.3 BrainWM-BAM

BrainWM-BAM (Figure 2c) is designed for latent action modeling, inspired by recent progress in world-action models (Schmidt & Jiang, 2024; Bruce et al., 2024). It infers fast and slow action tokens from brain-state transitions through prior–posterior modeling.

Prior–Posterior Dynamics Modeling. Given adjacent brain states z_t and z_{t+1} within a scan, an MLP posterior Q infers an action token capturing the observed transition, while an MLP prior P predicts it using optional generation conditions c_t available at time t:

$$a_t^{\text{post}} = Q(z_t,\ z_{t+1}), \qquad a_t^{\text{prior}} = P(c_t), \qquad c_t = \{m_t,\ e_{\text{text}},\ u_t\}, \qquad a_t \in \mathbb{R}^8, \tag{1}$$

where m_t summarizes the stimulus history up to time t, e_{text} embeds the text describing the subject and generation target, and u_t is the stimulus feature aligned to frame t. Conditions are randomly

dropped during training, allowing the prior to learn transition dynamics with or without these conditions. The generator G_θ, comprising the DiT and its sampling procedure, generates the next state from the current state and action token: $\hat{z}_{t+1} = G_\theta(z_t,\ a_t)$. The prior and posterior follow a teacher–student framework, as in latent action models (Schmidt & Jiang, 2024; Luo et al., 2026):

$$\mathcal{L}_{\text{align}} = \left\| a_t^{\text{prior}} - \text{sg}(a_t^{\text{post}}) \right\|^2, \qquad \mathcal{L}_{\text{gen}} = \left\| \hat{z}_{t+1} - z_{t+1} \right\|^2, \tag{2}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. $\mathcal{L}_{\text{gen}}$ denotes the generator objective in Section 3.4; at $s = 0$, the flow loss reduces to exactly this prediction error. The alignment loss trains the prior to match the posterior, while the posterior receives gradients only from the generation loss. The prior thus learns the transition itself, using conditions as cues when available. During training, the generator receives posterior tokens for a fraction p of each batch and prior tokens for the rest, with p decreasing linearly from 0.8 to 0.1. At inference, only the prior is used: $a_t = a_t^{\text{prior}}$. Appendix A details the anti-collapse constraints.

Fast and Slow Action Tokens. The per-frame token above is the fast action a_t, used mainly for within-run forecasting. For cross-run transitions, such as rest-to-task switches or aging between visits, BAM introduces a slow action $b \in \mathbb{R}^{32}$, learned by separate prior and posterior modules:

$$b^{\text{post}} = Q_s(S_A,\ S_B), \qquad b^{\text{prior}} = P_s(S_A,\ e_{\text{text}},\ \eta), \tag{3}$$

where S_A and S_B are the source and target scans, the text describes the target state, and η encodes the target age and visit interval for longitudinal simulation. Fast and slow actions are concatenated as $A_t = [a_t;\ b]$ and injected into each DiT frame through AdaLN via a zero-initialized MLP.

3.4 Multi-stage Training

Phase-I: Unified VAE Latent Pretraining. BrainWM-VAE reconstructs paired sMRI and fMRI with $\mathcal{L}_{\text{I}} = \mathcal{L}_s + \mathcal{L}_f + \beta\,\text{KL}$. Both modality-specific terms combine an L1 loss within the brain mask, a zero-background constraint, and a wavelet loss (Yasinzai et al., 2026); $\mathcal{L}_s$ adds an SSIM loss (Wang et al., 2004), and $\mathcal{L}_f$ adds four temporal losses to preserve fMRI dynamics (Appendix C).

Phase-II: Generative Forecasting Pretraining. The DiT and BAM are trained from scratch in the frozen latent space to generate the next window W_{k+1} from W_k, using flow matching (Lipman et al., 2023) with a stochastic interpolant (Albergo et al., 2023; Chen et al., 2024):

$$I_s = (1-s)\,W_k + s\,W_{k+1} + \epsilon(1-s)\sqrt{s}\,\xi, \qquad \mathcal{L}_{\text{flow}} = \mathbb{E}_{s,\xi}\left\| v_\theta(I_s, s, c) - \left(W_{k+1} - W_k - \epsilon\sqrt{s}\,\xi\right) \right\|^2, \tag{4}$$

where $s \in [0,1]$ is flow time, $\xi \sim \mathcal{N}(0, I)$, and ϵ is the noise scale. The model learns the drift of an SDE with diffusion coefficient $g(s) = \epsilon(1-s)$ rather than the corresponding probability-flow ODE velocity (Chen et al., 2024), and inference integrates that same SDE with Euler–Maruyama, adding noise $g(s)\sqrt{\Delta s}\,xi$ at every step but the last, so the sampler matches the training target. Empirically, SDE sampling preserves global variance and connectivity patterns but aligns poorly frame by frame, while single-step regression improves temporal correlation at the cost of variance. We therefore train both modes jointly, applying $\mathcal{L}_{\text{flow}}$ to all samples. For half the samples we draw $s \sim U(0,1]$ to learn the transport field across flow times; at inference, SDE integration generates the next window, which we assess by its spatial patterns across regions (*spatial generation*). For the other half we set $s = 0$ and the model predicts the next window in one step, $\hat{W}_{k+1} = W_k + v_\theta(W_k, 0, c)$, supervised additionally by the temporal correlation loss $\mathcal{L}_{\text{corr}} = 1 - \rho_t(\hat{W}_{k+1}, W_{k+1})$ (*temporal generation*). Both modes share the same velocity field and are trained with

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{flow}} + \mathbb{1}_{s=0}\,\lambda_c \mathcal{L}_{\text{corr}} + \mathbb{1}_{s\geq 0.5}\,\mathcal{L}_{\text{stat}} + \mathcal{L}_{\text{BAM}}, \tag{5}$$

where $\mathcal{L}_{\text{stat}}$ combines power-spectrum and spatial-covariance losses, and $\mathcal{L}_{\text{BAM}}$ includes the alignment and anti-collapse terms from Section 3.3 (Appendix C).

Phase-III: Generative Simulation Post-training. Starting from Phase-II weights, BrainWM is post-trained on cross-run source–target pairs for rest-to-task and longitudinal simulation, with slow-action tokens introduced at this stage. For rest-to-task, training models the transition from rest to the first task-state window, with the fast action set to zero so that the slow action alone encodes the

Table 1: Fine-tuning results on five downstream tasks. Bold: best overall; underline: best baseline. LA5c is external to BrainWM. Remaining tasks and linear-probe results are in Appendix E.1.

	ABIDE-I	ADNI	SRPBS		LA5c (external)
Model	Age regression RMSE↓ / r ↑	HC vs AD AUC↑ / F1↑	HC vs MDD AUC↑ / F1↑	HC vs SCZ AUC↑ / F1↑	HC vs BD AUC↑ / F1↑
BrainLM	5.180 / 0.688	0.617 / 0.512	0.733 / 0.625	0.766 / 0.642	0.425 / 0.471
Brain-JEPA	4.087 / 0.821	0.553 / 0.530	0.744 / 0.627	0.776 / 0.627	0.508 / 0.489
BrainMass	6.834 / 0.484	0.632 / 0.451	0.696 / 0.420	0.782 / 0.447	0.554 / 0.414
NeuroSTORM	3.902 / 0.845	0.640 / 0.525	0.677 / 0.551	0.749 / 0.550	0.531 / 0.454
Brain-DiT	4.286 / 0.803	0.594 / 0.550	0.577 / 0.448	0.630 / 0.489	0.465 / 0.353
BrainWM-VAE	3.997 / 0.832	**0.810** / 0.639	**0.751** / 0.644	0.814 / 0.597	0.475 / 0.454
BrainWM-DiT	**3.597** / **0.862**	0.775 / **0.673**	0.709 / 0.643	0.822 / **0.700**	**0.596** / 0.549
BrainWM-DiT+BAM	3.626 / 0.860	0.760 / 0.645	0.725 / **0.646**	**0.836** / 0.668	0.579 / **0.569**

switch. An additional identity amplitude term regresses each sample's per-voxel temporal standard deviation, targeting the subject-specific component overlooked by the main loss (Appendix C). For longitudinal simulation, pairs comprise two visits from the same subject, and the slow prior also receives the target age and visit interval.

4 Experiments

4.1 Setup

Pretraining data. Phases I and II used 59,915 paired sMRI–fMRI scans from 32 public cohorts, covering resting and task states in healthy participants and patients; Appendix B lists every cohort. We preprocessed all scans into MNI space using a common pipeline and spatially resampled them to 96^3, without temporal resampling. We split subjects 7:1:2, assigning all scans from each subject to the same split; LA5c and CHCP are held out entirely for external evaluation. We report all results on test splits.

Training and evaluation metrics. Phase I trained BrainWM-VAE on all paired training data and froze it at step 40k for latent extraction. Phase II trained the DiT and BAM from scratch on the 32 cohorts for 50 epochs using 6 H100 GPUs. Phase III performed two separate post-trainings from Phase-II weights: rest-to-task simulation on within-session HCP rest–task pairs, and longitudinal simulation on two-visit ADNI and PPMI pairs (Appendix C). During evaluation, we computed temporal correlation under *temporal generation* mode, and connectivity, spectral and variance metrics under *spatial generation* mode. Confidence intervals came from subject-level bootstrap resampling, and we tested paired subject-level differences against baselines by permutation (Appendix E).

4.2 Results-I: Representing the Brains

Setting. In fine-tuning, we trained an MLP task head on three BrainWM features: BrainWM-VAE uses the frozen VAE's latent mean μ_t; BrainWM-DiT uses the mean of the Phase-II DiT's intermediate tokens at $s = 0$; and BrainWM-DiT+BAM uses the latter with posterior action tokens injected. We zeroed the text and stimulus conditions during extraction to exclude label information from these inputs. ABIDE-I, ADNI, and SRPBS are internal to our pretraining corpus, while LA5c is external. We compared BrainWM with five fMRI foundation models (Appendices E.1 and C.4).

Results. Although BrainWM is not trained specifically for representation learning, its fine-tuned performance matched or exceeded that of dedicated foundation models (Table 1), including on external LA5c. Linear-probe results showed a consistent ranking (Appendix E.1). The VAE latent achieved the highest AUC on AD and MDD, with the largest margin on AD (0.81 versus 0.64 for the best baseline). We attribute this to structure–function fusion, given the anatomical information relevant to AD discrimination. Gains from the two dynamics-based features suggest that state transitions also carry discriminative information.

Table 2: Task-state forecasting under *temporal generation*: ROI temporal correlation R over the first 32, 64, and 128 generated frames, using each run's actual per-frame stimuli. CHCP is external; Appendix E.2 reports all 15 cohort–task settings.

		BrainWM			MONAI		
Cohort	Task	$h32$	$h64$	$h128$	$h32$	$h64$	$h128$
HCP	social	**0.348**	**0.241**	**0.155**	0.005	−0.001	0.013
HCP	relational	**0.146**	**0.179**	**0.111**	0.001	0.009	0.014
HCP	working memory	**0.175**	**0.093**	**0.097**	0.000	0.009	0.005
CHCP	social	**0.329**	**0.203**	**0.147**	0.006	−0.005	0.005
CHCP	relational	**0.193**	**0.216**	**0.103**	0.003	0.005	0.018
CHCP	working memory	**0.203**	**0.055**	**0.040**	−0.003	0.013	0.006

4.3 Results-II: Forecasting the Brains

Setting. We evaluated forecasting on both task and resting states after Phase-II pretraining. Task runs follow experimental designs, making their trajectories meaningful prediction targets, so we report ROI temporal correlation R under *temporal generation*; resting-state runs reflect spontaneous activity whose next frame is difficult to predict from history, so we assessed instead whether generated runs reproduce resting-state connectivity using *spatial generation*. Both modes generate 128 frames from one observed window, and task-state generation was conditioned on each run's own stimuli. As no comparable brain world model has public weights, we followed Seo et al. (2025) and Xia et al. (2026b) and trained a MONAI-based baseline (Appendix D.2).

Task states. Table 2 presents three tasks each from HCP and CHCP; the remaining task-state results are reported in Appendix E.2. BrainWM captured the observed trajectory at every horizon, whereas MONAI remained near chance level ($R \approx 0$), with similar performance on external CHCP. The advantage also held at the voxel level: BrainWM was closer to the observed fluctuations at 76.8% of brain voxels $\times$ frames (Figure 3). We also compared three training-free baselines: *repeat* tiles the observed window across the prediction horizon, *AR(1)* decays from the last observed frame toward the window mean, and *window mean* predicts that mean at every frame. Across 15 cohort–task settings, BrainWM led in all settings at $h32$, averaging 0.166 versus 0.043 for *AR(1)* and -0.043 for *repeat*. It also led in 110 of 120 setting-by-horizon comparisons against these two baselines. The negative correlation of *repeat* suggests that replaying the observed window is anticorrelated with subsequent activity. Paired subject-level comparisons showed gains of $+0.139$ over *repeat* and $+0.081$ over *AR(1)* at $h32$, with advantages over both at every horizon (328 subjects, $p < 0.0001$, permutation test; Appendix E.2).

Resting states. Across internal and external cohorts, BrainWM generated resting-state connectivity closer to observed scans than MONAI (Table 3): FC cosine was 0.68–0.76 versus 0.15–0.20, while a dFC cosine of 0.61–0.65 indicated preservation of dynamic connectivity patterns. MONAI came closest on PSD error, as BrainWM somewhat smoothed activity. Training-free baselines also left gaps. BrainWM exceeded *repeat* in FC cosine across all four cohorts, by $+0.016$ to $+0.080$ ($p < 0.01$). *AR(1)* decayed to nearly constant activity over 128 frames; on runs where its FC remained computable, its FC cosine did not differ significantly from BrainWM's. BrainWM, nevertheless, achieved substantially higher dFC cosine: 0.61–0.65 versus 0.16–0.30 for *AR(1)*. Its spectral error was also three to eight times smaller than that of either baseline. However, *repeat* better preserved the observed temporal variance (Appendix E.2).

4.4 Results-III: Simulating the Brains

Rest-to-task simulation. Given a resting run and a task description, we simulated the first 32-frame window of that subject's corresponding task scan. As shown in Table 4, a matched task description resulted in a correlation of 0.154 on HCP and 0.131 on the external CHCP. A mismatched task for the same subject cost 0.048 and 0.041, suggesting the model follows the specific task. Scanning the same person twice on the same task gave 0.134 between the two runs, so our generation was as close to the real one as a repeat scan. An average over other subjects doing the same task scored

Table 3: Resting-state forecasting under *spatial generation* over a 128-frame rollout: functional connectivity (FC) cosine, log-PSD mean absolute error, and dynamic FC (dFC) cosine. CHCP is external. Bold: better of the two.

	ABIDE-I		SRPBS		HCP rest		CHCP rest (ext.)	
Metric	BrainWM	MONAI	BrainWM	MONAI	BrainWM	MONAI	BrainWM	MONAI
FC cosine ↑	**0.746**	0.180	**0.755**	0.170	**0.676**	0.148	**0.692**	0.196
log-PSD MAE ↓	**1.838**	2.623	**1.944**	3.229	**1.892**	2.538	**2.045**	2.817
dFC cosine ↑	**0.609**	0.067	**0.654**	0.068	**0.618**	0.112	**0.647**	0.140

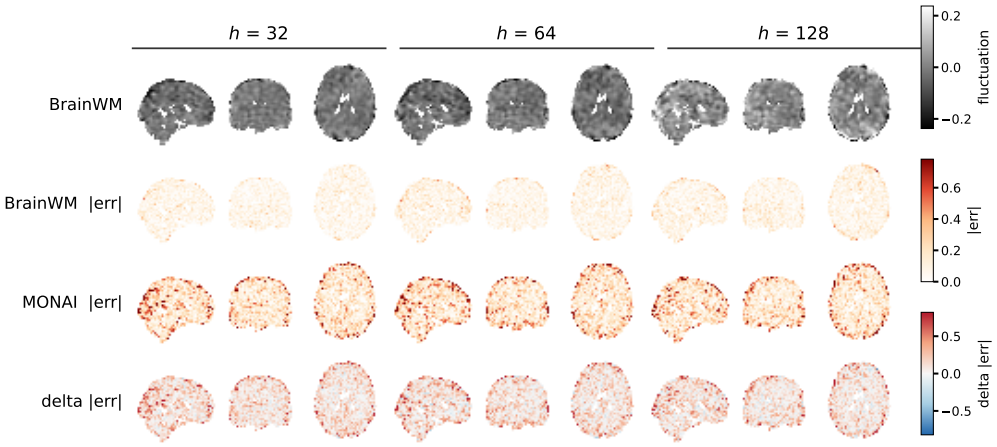

Figure 3: Voxel-level forecasting error on the 75th-percentile HCP working-memory run, at three rollout depths and anatomical planes. Rows: BrainWM's generated fluctuations, its absolute error, MONAI's absolute error, and the difference (red: BrainWM closer). BrainWM is closer at 76.8% of brain voxels × frames; mean absolute error 0.021 vs. 0.056.

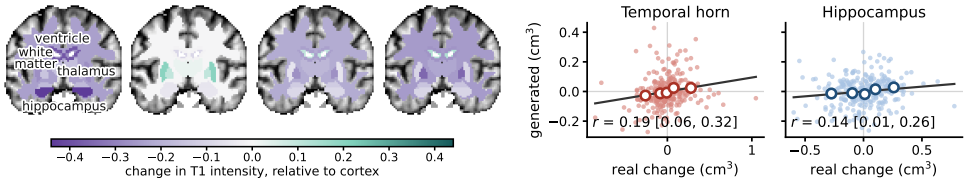

Figure 4: Longitudinal simulation. *Left*: regional T1 intensity change relative to each scan's cortex, under different elapsed intervals. *Right*: predicted versus observed volume change, adjusted for interval, cohort and baseline volume; dots are subjects, open circles quintile means.

temporal correlation higher (Appendix E.3), however, showing that individual-level dynamics are still limited in the temporal trajectories. Further assessment suggests that individual specificity lies in the spatial pattern, as indicated by subject-level fluctuation maps. Compared to task generation sourced from a subject's own resting scan, using another subject's resting scan while fixing the task description yielded negative similarity and a paired difference of 0.275 ($p < 0.001$, $n = 121$; Table 4). Directly using the source scan achieved comparable similarity, suggesting that generation preserves only source-subject information across conditions (Appendix E.3).

Longitudinal simulation. Given a baseline scan and an elapsed interval supplied to the slow action, we evaluated the model-generated follow-up structural scan. A zero interval produced almost no structural change, while doubling the actual interval produced greater change (Figure 4). Across 269 subjects, we compared predicted and observed volume changes in nine structures (Table 22), adjusting for interval, cohort, and baseline volume. The temporal horn ($r = 0.19$, $[0.06, 0.32]$) and the hippocampus ($r = 0.14$, $[0.01, 0.26]$) showed significant correlations. The two structures are

Table 4: Rest-to-task simulation against several references. *Matched*/*mismatched*: whether the task description is correct; *own*/*other*: whether the source resting scan is; *test*/*retest*: two real runs of one subject and task. Spatial correlation scores the 400-region deviation map of fluctuation amplitude. $^{**}p < 0.01$, $^{***}p < 0.001$.

Temporal correlation			*Spatial correlation (HCP, $n = 121$)*	
	HCP	CHCP (ext.)		HCP
matched vs. real	0.154	0.131	own vs. real	0.227
task effect (matched − mismatched)	$+0.048^{***}$	$+0.041^{**}$	subject effect (own − other)	$+0.275^{***}$
real test vs. real retest	0.134	0.185		

Table 5: BAM ablation. The no-BAM model is trained from scratch with the same budget and direct DiT conditioning. Values pool the 640 subjects both models scored across task and resting states, so they sit below the task-only averages of Table 2. Stars: paired differences from BrainWM, $^{*}p < 0.05$, $^{**}p < 0.01$ (Appendix F).

	$h32$	$h64$	$h96$	$h128$
BrainWM, *true*	0.081	0.055	0.046	0.042
BrainWM, *zero*	0.079	0.053	0.044	0.041
no BAM, *true*	0.066^{*}	0.049	0.044	0.042
no BAM, *zero*	0.058^{**}	0.039^{*}	0.033^{*}	0.032

closely relevant to Alzheimer's disease, which is associated with hippocampal atrophy and enlargement of the adjacent temporal horn. Cortical thinning, though measurable, was not predicted by our model (Appendix E.3).

4.5 Ablations

We trained a no-BAM variant from scratch under the same protocol, passing conditions directly to the DiT through AdaLN instead of action tokens. We evaluated both models on all task and resting-state test sets using either each run's actual per-frame stimuli (*true*) or an all-zero stimulus sequence (*zero*). BrainWM outperformed no-BAM with a significant advantage at $h32$ (Table 5). Removing stimuli had little effect on BrainWM but substantially reduced no-BAM performance, suggesting that BAM reduces reliance on stimulus inputs when modeling transitions. A third variant trained without the posterior was clearly worse than both at every horizon (Appendix F). Zeroing BrainWM's latent actions reduced performance by 0.005 at $h32$ and 0.012 at $h128$ ($p < 0.001$). Shifting the acquisition-time encoding by 48 TRs caused a drop of approximately 0.06 at every horizon, highlighting the importance of temporal alignment (Appendix F).

5 Conclusion and Limitations

We cast the brain as a dynamical system and introduce BrainWM, a multimodal world model for brain imaging. BrainWM represents brain states in a unified structure–function VAE latent space, models their evolution with a generative DiT, and learns transition-driving latent actions through BAM. Together they support representation learning, forecasting and conditional simulation: BrainWM matches or outperforms foundation models designed for representation while forecasting and simulating beyond their intended scope. To our knowledge, it is the first brain imaging world model evaluated across all three capabilities.

However, each capability can be improved. Representation quality largely depends on Phase-I training, where we observe competition between structure and function in the shared latent space. Forecasting is most accurate in the first generated block and degrades as the model consumes its own outputs. The generated-to-real variance ratio rises sharply over the four blocks, indicating exposure bias and making self-forcing post-training a possible next step (Huang et al., 2025). Rest-to-task simulation preserves individual spatial patterns but does not improve them, while longitudinal prediction is limited to two medial temporal structures rather than broader anatomical changes.

AI USE DISCLOSURE

We used generative AI tools to assist with code implementation and debugging, organize and interpret experimental results, help draft sections 3 and 4 of the manuscript, and polish the writing. We also used these tools to discuss possible explanations for observed results and suggest additional experiments. We tested the AI-assisted code, checked reported results against experimental outputs, and reviewed and revised the AI-assisted text. The authors take responsibility for the final manuscript, code, and reported findings.

ETHICS STATEMENT

This study uses existing neuroimaging datasets, described in Appendix B. Access to and reuse of these data are subject to the original providers' consent provisions, data-use agreements, and licensing terms. We do not use directly identifying personal information, such as names or contact details, as model inputs, nor do we include such information in the generated outputs.

BrainWM is a research model, not a clinical decision-making tool. Its generated trajectories are model predictions rather than observed outcomes. In particular, simulated task responses and longitudinal changes should not be interpreted as causal effects, individual prognoses, or evidence of treatment efficacy. Differences in participant demographics, acquisition protocols, and clinical populations may affect performance, and results from the evaluated cohorts may not generalize to underrepresented groups. Any clinical use would require separate validation and appropriate ethical and regulatory review.

REPRODUCIBILITY STATEMENT

Every component of the model is specified in Appendix A. Appendix C gives the three training phases with their losses and evaluation protocols. Appendix B lists all 34 registered cohorts with their subject and run counts, states the preprocessings each cohort took, and defines the subject-level splits and the training samples in each phases. Appendix D describes the trained generative baseline, the rules defining the untrained predictors, and the released protocol each downstream baseline follows. Appendix E reports every dataset-and-task setting behind the summary tables. The ablation models differ from the released BrainWM only in the configurations named in Appendix F. The source code is included as supplementary material.

REFERENCES

ADHD-200 Consortium. The ADHD-200 consortium: a model to advance the translational potential of neuroimaging in clinical neuroscience. *Frontiers in Systems Neuroscience*, 6:62, 2012.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega-Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, et al. An open resource for transdiagnostic research in pediatric mental health and learning disorders. *Scientific Data*, 4:170181, 2017.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Bharat B. Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M. Smith, Christian F. Beckmann, Jonathan S. Adelstein, Randy L. Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal Behbahani, Stephanie Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair,

Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

Josue Ortega Caro, Antonio H. de Oliveira Fonseca, Syed A. Rizvi, Matteo Rosati, Christopher Averill, James L. Cross, Prateek Mittal, Emanuele Zappala, Rahul M. Dhodapkar, Chadi G. Abdallah, and David van Dijk. BrainLM: A foundation model for brain activity recordings. In *International Conference on Learning Representations (ICLR)*, 2024.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes. In *International Conference on Machine Learning (ICML)*, 2024.

Greig I. de Zubicaray, Ming-Chang Chiang, Katie L. McMahon, David W. Shattuck, Arthur W. Toga, Nicholas G. Martin, Margaret J. Wright, and Paul M. Thompson. Meeting the challenges of neuroimaging genetics. *Brain Imaging and Behavior*, 2(4):258–263, 2008.

Adriana Di Martino, Chao-Gan Yan, Qingyang Li, Erin Denio, Francisco X. Castellanos, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Susan Y. Bookheimer, Mirella Dapretto, et al. The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659–667, 2014.

Adriana Di Martino, David O'Connor, Bosi Chen, Kaat Alaerts, Jeffrey S. Anderson, Michal Assaf, Joshua H. Balsters, Leslie Baxter, Anita Beggiato, Sylvie Bernaerts, et al. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Scientific Data*, 4: 170010, 2017.

Zijian Dong, Ruilin Li, Yilei Wu, Thuan Tinh Nguyen, Joanna Su Xian Chong, Fang Ji, Nathanael Ren Jie Tong, Christopher Li Hsian Chen, and Juan Helen Zhou. Brain-JEPA: Brain dynamics foundation model with gradient positioning and spatiotemporal masking. In *Advances in Neural Information Processing Systems*, 2024.

Zijian Dong, Ruilin Li, Joanna Su Xian Chong, Niousha Dehestani, Yinghui Teng, Yi Lin, Zhizhou Li, Yichi Zhang, Yapei Xie, Leon Qi Rong Ooi, B. T. Thomas Yeo, and Juan Helen Zhou. Brain harmony: A multimodal foundation model unifying morphology and function into 1D tokens. *arXiv preprint*, 2025.

Zijian Dong, Jianxiong Zhou, Kwun Kei Ng, Jan Paolo Macapinlac Balagtas, Zhizhou Li, Zijiao Chen, and Juan Helen Zhou. NeuroWorld: A latent brain world model for stimulus-conditioned human brain dynamics. *arXiv preprint*, 2026.

Peiyu Duan, Jiyao Wang, Nicha C. Dvornek, Junlin Yang, Ziqi Gao, Lawrence H. Staib, and James S. Duncan. FM-fMRI: Event conditioned flow matching for rest-to-task fMRI time-series synthesis. *arXiv preprint arXiv:2605.26423*, 2026.

Karl J. Friston, Paul Fletcher, Oliver Josephs, Andrew Holmes, Michael D. Rugg, and Robert Turner. Event-related fMRI: characterizing differential responses. *NeuroImage*, 7(1):30–40, 1998.

Peng Gao, Hao-Ming Dong, Si-Man Liu, Xue-Ru Fan, Chao Jiang, Yin-Shan Wang, Daniel Margulies, Hui-Fang Li, and Xi-Nian Zuo. A Chinese multi-modal neuroimaging data release for increasing diversity of human brain mapping. *Scientific Data*, 9:286, 2022.

Jianqiao Ge, Guoyuan Yang, Meizhen Han, Suyu Zhou, Weiwei Men, Lang Qin, Bingjiang Lyu, Hai Li, Haobo Wang, Hengyi Rao, et al. Increasing diversity in connectomics with the Chinese Human Connectome Project. *Nature Neuroscience*, 26(1):163–172, 2023.

Salah Ghamizi, Georgia Kanli, Yu Deng, V. Palissot, Magali Perquin, and Olivier Keunen. Foundation models for brain imaging: A systematic review. *NeuroImage*, 333:121877, 2026. doi: 10.1016/j.neuroimage.2026.121877.

Avram J. Holmes, Marisa O. Hollinshead, Timothy M. O'Keefe, Victor I. Petrov, Gabriele R. Fariello, Lawrence L. Wald, Bruce Fischl, Bruce R. Rosen, Ross W. Mair, Joshua L. Roffman, et al. Brain Genomics Superstruct Project initial data release with structural, functional, and behavioral measures. *Scientific Data*, 2:150031, 2015.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *arXiv preprint*, 2025.

Clifford R. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chadwick Ward, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Brian P. Keane, Deanna M. Barch, Ravi D. Mill, Steven M. Silverstein, Bart Krekelberg, and Michael W. Cole. Brain network mechanisms of visual shape completion. *NeuroImage*, 236: 118069, 2021.

Clare Kelly, Xi-Nian Zuo, Kristin Gotimer, Christine L. Cox, Lauren Lynch, Davide Brock, Davide Imperati, Hugh Garavan, John Rotrosen, F. Xavier Castellanos, and Michael P. Milham. Reduced interhemispheric resting state functional connectivity in cocaine addiction. *Biological Psychiatry*, 69(7):684–692, 2011.

Kristen M. Kennedy, Karen M. Rodrigue, Gerard N. Bischof, Andrew C. Hebrank, Patricia A. Reuter-Lorenz, and Denise C. Park. Age trajectories of functional activation under conditions of low and high processing demands: an adult lifespan fMRI study. *NeuroImage*, 104:21–34, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Dorit Kliemann, Ralph Adolphs, Tim Armstrong, Paola Galdi, David A. Kahn, Tessa Rusch, A. Zeynep Enkavi, Deuhua Liang, Steven Lograsso, Wenying Zhu, et al. Caltech Conte Center, a multimodal data resource for exploring social cognition and decision-making. *Scientific Data*, 9:138, 2022.

Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. *arXiv preprint arXiv:2412.08802*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Siman Liu, Yin-Shan Wang, Qing Zhang, Quan Zhou, Li-Zhi Cao, Chao Jiang, Zhe Zhang, Ning Yang, Qi Dong, and Xi-Nian Zuo. Chinese Color Nest Project: An accelerated longitudinal brain-mind cohort. *Developmental Cognitive Neuroscience*, 52:101020, 2021.

Wei Liu, Dongtao Wei, Qunlin Chen, Wenjing Yang, Jie Meng, Guorong Wu, Taiyong Bi, Qinglin Zhang, Xi-Nian Zuo, and Jiang Qiu. Longitudinal test-retest neuroimaging data from healthy young adults in southwest China. *Scientific Data*, 4:170017, 2017.

Hao Luo et al. Being-H0.7: A latent world-action model from egocentric videos. *arXiv preprint*, 2026.

Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Christopher Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The Parkinson Progression Marker Initiative (PPMI). *Progress in Neurobiology*, 95(4):629–635, 2011.

Kate B. Nooner, Stanley J. Colcombe, Russell H. Tobe, Maarten Mennes, Melissa M. Benedict, Alexis L. Moreno, Laura J. Panek, Shaquanna Brown, Stephen T. Zavitz, Qingyang Li, et al. The NKI-Rockland sample: a model for accelerating the pace of discovery science in psychiatry. *Frontiers in Neuroscience*, 6:152, 2012.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. da Costa, Ashay Patel, et al. Generative AI for medical imaging: Extending the MONAI framework. *arXiv preprint arXiv:2307.15208*, 2023.

Russell A. Poldrack, Eliza Congdon, William Triplett, Krzysztof J. Gorgolewski, Katherine H. Karlsgodt, Jeanette A. Mumford, Fred W. Sabb, Nelson B. Freimer, Edythe D. London, Tyrone D. Cannon, and Robert M. Bilder. A phenome-wide examination of neural and cognitive function. *Scientific Data*, 3:160110, 2016.

Chris Racey, Christina Kampoureli, Oscar Bowen-Hill, Mathilde Bauer, Ivor Simpson, Charlotte Rae, Magda del Rio, Julia Simner, and Jamie Ward. An open science MRI database of over 100 synaesthetic brains and accompanying deep phenotypic information. *Scientific Data*, 10:766, 2023.

Jianxun Ren, Ning An, Cong Lin, Youjia Zhang, Zhenyu Sun, Wei Zhang, Shiyi Li, Ning Guo, Weigang Cui, Qingyu Hu, Weiwei Wang, Xuehai Wu, Yong Wang, Tao Jiang, Theodore D. Satterthwaite, Danhong Wang, and Hesheng Liu. DeepPrep: an accelerated, scalable and robust pipeline for neuroimaging preprocessing empowered by deep learning. *Nature Methods*, 22:473–484, 2025.

Giovanni A. Salum, Ary Gadelha, Pedro M. Pan, Tais S. Moriyama, Ana Soledade Graeff-Martins, Ana Carina Tamanaha, Pedro Alvarenga, Fernanda Valle Krieger, Bacy Fleitlich-Bilyk, Andrea Jackowski, et al. High risk cohort study for psychiatric disorders in childhood: rationale, design, methods and preliminary results. *International Journal of Methods in Psychiatric Research*, 24 (1):58–73, 2015.

Theodore D. Satterthwaite, Mark A. Elliott, Kosha Ruparel, James Loughead, Karthik Prabhakaran, Monica E. Calkins, Ryan Hopson, Chad Jackson, Jack Keefe, Marisa Riley, et al. Neuroimaging of the Philadelphia Neurodevelopmental Cohort. *NeuroImage*, 86:544–553, 2014.

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.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *International Conference on Learning Representations*, 2024.

Jungwoo Seo, David Keetae Park, Shinjae Yoo, and Jiook Cha. Scalable diffusion transformer for conditional 4D fMRI synthesis. In *NeurIPS Workshop*, 2025.

Lukas Snoek, Maite M. van der Miesen, Tinka Beemsterboer, Andries van der Leij, Annemarie Eigenhuis, and H. Steven Scholte. The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8:85, 2021.

R. Nathan Spreng, Roni Setton, Udi Alter, Benjamin N. Cassidy, Bri Darboh, Elizabeth DuPre, Karin Kantarovich, Amber W. Lockrow, Laetitia Mwilambwe-Tshilobo, Wen-Ming Luh, Prantik Kundu, and Gary R. Turner. Neurocognitive aging data release with behavioral, structural and multi-echo functional MRI measures. *Scientific Data*, 9:119, 2022.

Lachlan T. Strike, Narelle K. Hansell, Kai-Hsiang Chuang, Jessica L. Miller, Greig I. de Zubicaray, Paul M. Thompson, Katie L. McMahon, and Margaret J. Wright. Queensland Twin Adolescent Brain project, a longitudinal study of adolescent brain development. *Scientific Data*, 10:195, 2023.

Divyanshu Tak et al. A generalizable foundation model for analysis of human brain MRI. *Nature Neuroscience*, 2026.

Saori C. Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, 8:227, 2021.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E. J. Behrens, Essa Yacoub, and Kamil Ugurbil. The WU-Minn Human Connectome Project: an overview. *NeuroImage*, 80: 62–79, 2013.

Cheng Wang, Yu Jiang, Zhihao Peng, Chenxin Li, Chang-bae Bang, Lin Zhao, Wanyi Fu, Jinglei Lv, Jorge Sepulcre, Carl Yang, Lifang He, Tianming Liu, Xue-Jun Kong, Quanzheng Li, Daniel S. Barron, Anqi Qiu, Randy Hirschtick, Byung-Hoon Kim, Hongbin Han, Xiang Li, and Yixuan Yuan. Towards a general-purpose foundation model for functional MRI analysis. *Nature Biomedical Engineering*, 2026a. doi: 10.1038/s41551-026-01666-y.

Mo Wang, Junfeng Xia, Wenhao Ye, Enyu Liu, Kaining Peng, Jianfeng Feng, Quanying Liu, and Hongkai Wen. SLIM-Brain: A data- and training-efficient foundation model for fMRI data analysis. *arXiv preprint*, 2025.

Mo Wang, Wenhao Ye, Junfeng Xia, Junxiang Zhang, Xuanye Pan, Minghao Xu, Haotian Deng, Hongkai Wen, and Quanying Liu. Omni-fMRI: A universal atlas-free fMRI foundation model. *arXiv preprint*, 2026b.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Dongtao Wei, Kaixiang Zhuang, Lei Ai, Qunlin Chen, Wenjing Yang, Wei Liu, Kangcheng Wang, Jiangzhou Sun, and Jiang Qiu. Structural and functional brain scans from the cross-sectional Southwest University Adult Lifespan Dataset. *Scientific Data*, 5:180134, 2018.

Junfeng Xia, Wenhao Ye, Xuanye Pan, Xinke Shen, Mo Wang, and Quanying Liu. Brain-DiT: A universal multi-state fMRI foundation model with metadata-conditioned pretraining. *arXiv preprint arXiv:2604.12683*, 2026a.

Junfeng Xia, Wenhao Ye, Junxiang Zhang, Xuanye Pan, Mo Wang, and Quanying Liu. BrainWorld: A structural-prior-conditioned generative model for whole-brain 4D fMRI dynamics. *arXiv preprint arXiv:2606.17742*, 2026b.

Yanwu Yang, Chenfei Ye, Guinan Su, Ziyao Zhang, Zhikai Chang, Hairui Chen, Piu Chan, Yue Yu, and Ting Ma. BrainMass: Advancing brain network analysis for diagnosis with large-scale self-supervised learning. *IEEE Transactions on Medical Imaging*, 2024.

Muhammad Nabi Yasinzai, Remika Mito, and Mangor Pedersen. Wavelet-fusion diffusion model for multimodal brain MRI synthesis with modality and metadata conditioning. *arXiv preprint*, 2026.

Xi-Nian Zuo, Jeffrey S. Anderson, Pierre Bellec, Rasmus M. Birn, Bharat B. Biswal, Janusch Blautzik, John C. S. Breitner, Randy L. Buckner, Vince D. Calhoun, F. Xavier Castellanos, et al. An open science resource for establishing reliability and reproducibility in functional connectomics. *Scientific Data*, 1:140049, 2014.

A Additional Details of BrainWM

This appendix provides implementation details for Sections 3.1 to 3.3. We define the notation, describe the architecture and conditioning inputs, and explain the normalization and training constraints used in BAM. We also report parameter counts for each module.

A.1 Notation

Window-level implementation. Section 3.3 introduces the prior and posterior using adjacent states z_t and z_{t+1}. In the implementation, transitions are modeled between 32-frame windows. The posterior compares spatially pooled features at corresponding frame positions in W_k and W_{k+1} ($\bar{w}_t \in \mathbb{R}^{32}$) and produces one action token per frame: $a_t^{\text{post}} = Q(\bar{w}_t^k, \bar{w}_t^{k+1})$.

Table 6: Notation used in the main text and this appendix.

Symbol	Meaning
x^s, x_t^f	Structural image (96^3) and the t-th functional frame (96^3).
$z_t = \mu_t \in \mathbb{R}^{32\times 12^3}$	Unified structure–function latent of frame t.
$W_k = z_{32k+1:32k+32}$	The k-th 32-frame latent window, the generation unit of the DiT.
I_s, s	Intermediate state of the flow from W_k to W_{k+1} and its flow time, $s \in [0, 1]$.
$v_\theta(I_s, s, c)$	Velocity field predicted by the DiT; c denotes all conditions.
G_θ	Generator: DiT and the sampling process.
$a_t \in \mathbb{R}^8$	Fast action (per frame); a_t^{post} / a_t^{prior} are its posterior / prior versions.
$b \in \mathbb{R}^{32}$	Slow action (one per run); b^{post} / b^{prior}.
$A_t = [a_t;\, b]$, $g_t = G(A_t)$	Action tokens and their per-frame injection into the DiT through a zero-initialized MLP.
Q, P; Q_s, P_s	Posterior and prior modules of the fast / slow action.
m_t, e_{text}, u_t	Memory summary of the stimulus history, text condition embedding, per-frame stimulus feature.
S_A, S_B, η	Segment summaries of the source / target scan; age-interval feature.
ρ	Fraction of samples in a batch fixed at $s = 0$ (Phase II 0.5, rest-to-task 1, longitudinal 0).
$\mathcal{L}_{\text{gen}}$	The schematic generation loss of Section 3.3; in practice the Phase-II losses acting on the generator, $\mathcal{L}_{\text{flow}} + \mathbb{1}_{s=0}(\lambda_c \mathcal{L}_{\text{corr}} + \lambda_m \mathcal{L}_{\text{mean}})$, from which alone the posterior Q receives gradients.
λ_c, $\mathcal{L}_{\text{stat}}$	Weight of the temporal correlation loss (1.0); power-spectrum and spatial-covariance losses on high-s samples (0.05 each).
$\mathcal{L}_{\text{BAM}}$	Alignment loss + variance floor + swap margin (Appendix A.5).

A.2 BRAINWM-VAE

- **Architecture.** Two modality-specific 3D convolutional encoders map 96^3 volumes to a 12^3 grid through four residual stages. Their outputs are concatenated, and a fusion head produces μ and $\log \sigma^2$ with 32 channels. Two modality-specific decoders mirror the encoders. Both observed and generated latents can therefore be decoded into structural or functional images. In longitudinal simulation, the structural decoder reconstructs a future T1 image from the generated future state (Appendix C.3).
- **Gradient isolation.** Cross-modal gradients between the two encoders are blocked: the structural reconstruction loss does not update the functional encoder, and vice versa. The fusion head and both decoders are trained by both losses. This prevents either modality's loss from changing the other modality's encoder.
- **Latent convention.** Throughout the main text, z_t denotes the posterior mean μ_t. During Phase I training, we sample $z = \mu + \sigma\epsilon$ from the posterior. Phases II and III and all downstream evaluation use μ_t only. Before the latents enter the world model, a single scalar estimated on the corpus rescales them to unit variance.

A.3 BRAINWM-DIT

- **Tokens and blocks.** Each 3^3 patch is linearly projected to 768 dimensions, yielding 64 tokens per frame and 2048 per 32-frame window. Twelve transformer blocks apply self-attention across all spatial positions and frames in the window. No causal mask is used, since the model generates a window rather than an individual frame.
- **Conditioning.** Each DiT block reads eight text tokens through cross-attention. These tokens are projected from the joint embedding e_{text} produced by jina-clip-v2 (Koukounas et al., 2024) from the run metadata and condition file. In Phase II, the entire text embedding is dropped with probability 0.30. Frame-specific conditioning is supplied through the AdaLN modulation vector

$$c_t = \text{LN}_{\text{no-affine}}\big(\text{MLP}([\bar{e};\ \psi(s)]) + g_t\big), \tag{6}$$

where $\bar{e}$ is an attention-pooled summary of the condition tokens, $\psi(s)$ is the flow-time embedding, and g_t is the BAM action feature. The same vector modulates every block, and the outer, parameter-free LayerNorm holds its scale fixed.

- **The generator G_θ.** The DiT takes I_s together with spatial, acquisition-time, and flow-time embeddings and predicts the velocity field $v_\theta(I_s, s, c)$. The generator in Section 3.3 comprises the DiT and its sampling procedure; it introduces no additional parameters. Temporal generation uses a single step at $s = 0$, $\hat{W}_{k+1} = W_k + v_\theta(W_k, 0, c)$. Spatial generation integrates the SDE in Section 3.2 from W_k to $s = 1$ using 50 Euler–Maruyama steps.

A.4 Positional Embeddings

We add three embeddings to the tokens. A fixed 3D sinusoidal embedding encodes position on the spatial token grid. A sinusoidal embedding of flow time s is also included in the modulation vector c_t. The physical-time embedding from Section 3.2 is

$$\phi_t = \mathrm{MLP}\Big(\big[\, i_t/10^3,\ \tau_t/300,\ \log(\mathrm{TR}/0.72),\ \{\sin,\cos\}(2\pi\tau_t/P_k)\big]_{k=1}^{7}\Big), \tag{7}$$

where i_t is the frame index, τ_t is the acquisition time in seconds from the start of the run, and $P_k \in \{4, 8, \ldots, 256\}$ s are the seven Fourier periods. This embedding is added to every token in the frame. The 32 cohorts have different sampling rates, so a given frame index does not correspond to the same elapsed time across runs. Including τ_t and TR allows the model to retain each run's acquisition timing.

A.5 BrainWM-BAM

- **Fast posterior Q.** For each frame position, the posterior takes the spatially pooled features of the two windows and their difference. Providing the difference explicitly encourages the action to describe change. With only the two states as input, the token instead degenerates into a subject-specific representation that remains constant within a run.
- **Fast prior P.** Memory, text, and stimulus inputs have separate projection pathways. Their outputs are concatenated and fused into an action token. The stimulus pathway uses no normalization, since its per-frame magnitude carries the haemodynamic response in block-design tasks; LayerNorm would turn this gradual response into a square wave. For movies, visual and auditory tracks are encoded separately and summed. We also omit normalization after concatenation to avoid rescaling weaker pathways according to the statistics of stronger ones. Each pathway is dropped with probability 0.15 per sample, allowing the prior to operate with missing conditions.
- **Limiting action capacity.** The action has eight dimensions, and noise ($\sigma = 0.5$) is added to the posterior token before the generator sees it, while alignment uses the clean token. The posterior has access to the target window, whereas the prior does not. We limit the action token's capacity so that it describes the condition and state change rather than encoding the target window directly.
- **Mixing and annealing.** Each batch contains samples conditioned on posterior tokens and samples conditioned on prior tokens. The posterior fraction decreases linearly from 0.8 to 0.1 over 16k steps, gradually shifting training toward the prior tokens used at inference.
- **Anti-collapse constraints** (the "additional constraints" of Section 3.3):

$$\mathcal{L}_{\mathrm{BAM}} = 1.0 \cdot \mathcal{L}_{\mathrm{align}} + 1.0 \cdot \sum_d \mathrm{relu}\big(0.5 - \mathrm{std}_B(a_{\cdot d}^{\mathrm{post}})\big) + 0.5 \cdot \mathcal{L}_{\mathrm{swap}}. \tag{8}$$

  The second term places a lower bound on the within-batch standard deviation of each action dimension, following VICReg (Bardes et al., 2022), to prevent the posterior from collapsing to a constant. The swap loss, $\mathcal{L}_{\mathrm{swap}} = \mathrm{relu}\big(0.02 + \mathcal{L}_{\mathrm{gen}}(a) - \mathcal{L}_{\mathrm{gen}}(\tilde{a})\big)$, uses another sample's action token $\tilde{a}$. It penalizes the generator unless the correct token yields a lower generation loss than the swapped token, encouraging the generator to use the action.
- **Implementation of $\mathcal{L}_{\mathrm{gen}}$.** Section 3.3 uses $\|\hat{z}_{t+1} - z_{t+1}\|^2$ as a schematic generation loss. In practice, the generator is trained with the Phase-II losses: $\mathcal{L}_{\mathrm{flow}}$ for samples with $s > 0$, and $\lambda_c \mathcal{L}_{\mathrm{corr}} + \lambda_m \mathcal{L}_{\mathrm{mean}}$ for samples with $s = 0$. The posterior Q receives gradients only through these losses on samples conditioned on posterior tokens.

- **Slow action.** The slow posterior takes summaries of the source and target runs. The slow prior takes the source summary, text, and age interval. Each run is summarized by four segment means rather than a single run mean, preserving information about how the state varies over the scan. Alignment, mixing, and annealing follow the fast-action procedure. Longitudinal changes can have similar directions and different magnitudes: ageing over one year should not be represented in the same way as ageing over nine years. We therefore use

$$b = \mathrm{LN}(\hat{b}) \cdot \mathrm{softplus}(\kappa), \tag{9}$$

  where the posterior magnitude is given by $\|\bar{z}_B - \bar{z}_A\|$ and the prior magnitude is predicted from the age interval. The two are aligned on a log scale. This magnitude scaling is enabled only during longitudinal post-training. In Phase II, all samples share a learnable slow-token placeholder; the slow posterior and prior are not trained. These modules are activated in Phase III. With zero-initialized injection and the existing placeholder, Phase III starts from the Phase-II model without changing its initial behavior. For the switching window in rest-to-task simulation, the fast action is set to zero so that the slow action alone represents the state change.
- **Injection** G**.** A zero-initialized, two-layer MLP projects the concatenated fast and slow tokens into the per-frame modulation vector c_t. The slow token is broadcast across frames. Zero initialization preserves the base model's output at the start of fine-tuning.

A.6 Prior Conditions

Table 7: Conditions read by the fast and slow priors.

Prior	Condition	Source	Dim.	Norm.	Phase
Fast	Memory m_t	Pooled stimulus tracks of the past 3 windows	1536→64	none (magnitude is the signal)	II/III
Fast	Text e_{text}	jina-clip-v2 joint embedding of run metadata + condition JSON	768	LN	II/III
Fast	Stimulus track u_t	Condition-sentence embedding × event mask × SPM double-gamma HRF (sampled at the run's TR)	768 (task) / 1024+512 (movie)	none	II/III (task cohorts)
Slow	Source summary S_A	Source-run latents, uniformly subsampled to $\leq$256 frames, 4-segment pooling	128	LN	III
Slow	Text	As above (target condition sentence)	768	LN	III
Slow	Age interval η	Target age and Δmonths → 8-octave Fourier features + raw values	34	none	III (longitudinal)

The entire text embedding is dropped with probability 0.30 in Phase II, as is the slow prior's text input in Phase III. Stimulus dropout is 0.15 in Phase II and 0.25 in rest-to-task post-training. Longitudinal pairs and resting cohorts have no stimulus track, represented by a zero vector with `present = false`. We describe the conditioning inputs below.

Text. Each run pair has a JSON condition file containing source and target scan metadata and a sentence describing the target. The sentence describes the participant's clinical background and task, without information derived from the images. There is no separate diagnosis field; clinical information is supplied only through this description.

All three phases use the same template. In Phase II, source and target metadata refer to the same run. In rest-to-task simulation, they refer to a resting run and a task run, respectively. The target sentence is identical to the one used for that task run in Phase II; the source–target metadata specify

the transition. Longitudinal pairs add two clauses giving the participant's age at the target scan and the elapsed time since the source. Figure 5 shows examples of all three cases. A Phase-II task pair differs from the resting example only in the state reported in both metadata blocks. jina-clip-v2 (Koukounas et al., 2024) jointly embeds the metadata and sentence. Both priors and the DiT's cross-attention use this embedding.

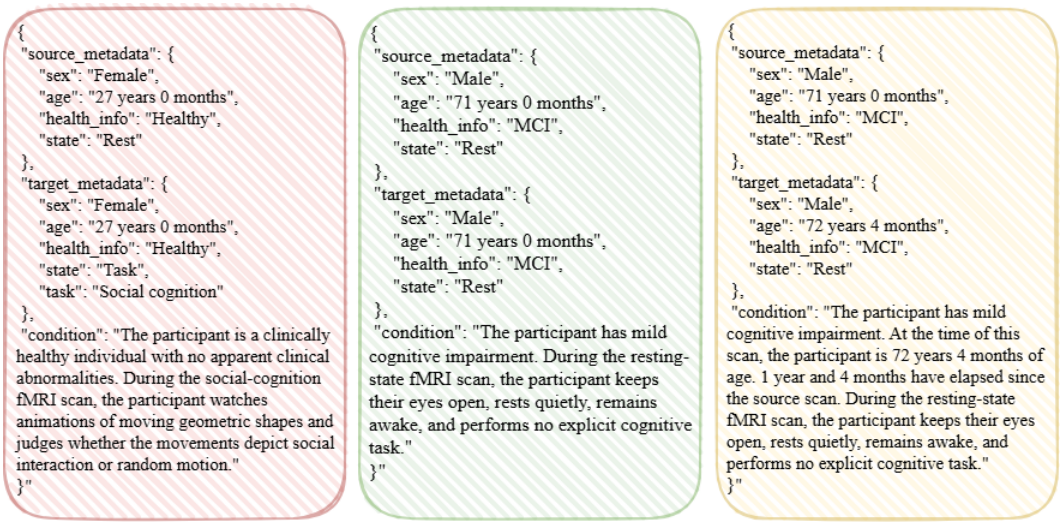

Figure 5: Example condition files. *Left*, rest-to-task: the target sentence describes the task run, while the source metadata specify a resting scan. *Middle*, Phase II: source and target refer to the same run. *Right*, longitudinal: two clauses are added to the middle example to specify the age at the target scan and the elapsed interval.

Stimulus track. The stimulus track provides frame-level information about the experimental paradigm. For cohorts with event files, we embed a sentence for each condition with jina-clip-v2, activate it over the blocks the FSL three-column event file assigns to that condition, and convolve the sequence with the canonical haemodynamic response (Friston et al., 1998). This accounts for the delay and gradual rise and fall of the blood-flow response relative to the events. HCP onsets are shifted by each task's fixed offset from the start of the scan.

In AOMIC-ID1000, all participants watch the same film from beginning to end, so the stimulus is represented by the film rather than a list of task events. Visual and auditory features at each time point are embedded and passed through the same haemodynamic filter. Event-based and movie-based tracks do not occur within the same run and use separate encoders in the prior. Resting runs and longitudinal pairs have no stimulus track.

Memory. Memory is the stimulus track averaged over the preceding three windows, roughly the preceding minute of scanning. It supplies recent stimulus levels that are not captured by the current frame or the change-oriented fast action. We leave this summary unnormalized to preserve its magnitude. Memory is computed from the stimulus rather than generated latents, keeping it tied to the observed stimulus history and avoiding feedback from generation errors during long rollouts.

Age interval. The slow action uses the participant's age at the target scan and the months elapsed since the source scan to specify the extent of ageing. Both values are encoded with sine and cosine features at multiple frequencies, with the raw values appended. The raw values distinguish intervals that would otherwise have identical periodic embeddings. Without them, a ten-year interval produced the same input as a zero interval, making it ambiguous which follow-up interval is being evaluated in Appendix C.3. If age is unavailable for a scan, its age input is set to zero.

A.7 Parameter Counts

Table 8 lists the parameter counts for each module. The VAE is trained in Phase I and frozen thereafter; only the world model is optimized in Phases II and III.

Table 8: Parameters per module. The two encoders and the two decoders are identical in size, one of each per modality.

Stage	Module	Parameters
BrainWM-VAE	Encoder (each)	1.82M
	Fusion head	2.69M
	Decoder (each)	1.82M
	Total	**9.96M**
World model	BrainWM-DiT	175.0M
	BAM fast posterior	0.03M
	BAM fast prior	1.33M
	BAM slow module	0.45M
	Action injection and projections	1.9M
	Total trainable	**178.7M**

B Datasets

Dataset overview. Table 9 summarizes all registered cohorts, including the numbers of subjects, T1 scans, and resting-state and task fMRI runs, along with the train/validation/test splits. The final three columns indicate whether each cohort is used in the corresponding training phase. The 32 cohorts marked in column II form the pretraining set, containing 59,915 fMRI runs paired with sMRI. CHCP and LA5c are reserved for external evaluation and are not used in any training phase. The cohort sources include: ABIDE-I (Di Martino et al., 2014), ABIDE-II (Di Martino et al., 2017), ACPI (Kelly et al., 2011), ADHD-200 (ADHD-200 Consortium, 2012), ADNI (Jack et al., 2008), AOMIC (Snoek et al., 2021), the Beijing cohorts and FCP-1000 (Biswal et al., 2010), BHRC (Salum et al., 2015), Caltech Conte (Kliemann et al., 2022), CCNP (Liu et al., 2021), CHCP (Ge et al., 2023), CoRR (Zuo et al., 2014), DLBS (Kennedy et al., 2015), ds003404 (Keane et al., 2021), GSP (Holmes et al., 2015), HBN (Alexander et al., 2017), HCP (Van Essen et al., 2013), ISYB (Gao et al., 2022), LA5c (Poldrack et al., 2016), NAD (Spreng et al., 2022), NKI (Nooner et al., 2012), PNC (Satterthwaite et al., 2014), PPMI (Marek et al., 2011), QTAB (Strike et al., 2023), QTIM (de Zubicaray et al., 2008), SALD (Wei et al., 2018), SLIM (Liu et al., 2017), SRPBS (Tanaka et al., 2021), and Synaesthesia (Racey et al., 2023).

Preprocessing. We preprocessed most cohorts with DeepPrep (Ren et al., 2025). For ABIDE-I, ABIDE-II, ADHD-200, the three AOMIC releases, and Caltech Conte Center, we used the publicly released preprocessed derivatives and brought them into MNI152NLin6Asym space at 2 mm. Both routes then follow the same processing steps: resampling to 96^3, discarding the first ten volumes of each fMRI run to allow the signal to stabilize, z-scoring within the brain mask, and storing the data as float16. We sampled 32-frame windows when preparing model inputs. Each run retains its native TR, which is supplied to the physical-time embedding rather than resampling the time series to a common rate (Appendix A.4). For GSP, HCP, CHCP, ADNI, and PPMI, the table reports the subsets we processed, not the full cohorts. Time constraints prevented us from processing these cohorts in full. The included subsets consist of the data processed before we stopped, with no additional selection criteria.

Splits. We divide subjects into training, validation, and test sets in a 7:1:2 ratio. All runs and transition pairs from a subject remain in the same split. These assignments are fixed throughout training and evaluation.

Phase-III transition pairs. Rest-to-task pairs use resting and task runs from the same session with matching phase-encoding directions. HCP provides the training and internal test pairs, while CHCP

Table 9: Dataset overview. Columns I, II, and III indicate use in each training phase.

Cohort	Subj.	T1	fMRI	Rest	Task	Train/Val/Test	I	II	III
ABIDE-I	987	987	1036	1036	0	720/107/209	Y	Y	
ABIDE-II	586	586	778	778	0	538/83/157	Y	Y	
ACPI	163	163	196	196	0	138/19/39	Y	Y	
ADHD-200	598	598	863	818	45	613/81/169	Y	Y	
ADNI	1585	3911	3911	3911	0	2748/378/785	Y	Y	Y
AOMIC-ID1000	881	881	881	0	881	617/88/176	Y	Y	
AOMIC-PIOP1	216	216	1236	0	1236	881/114/241	Y	Y	Y
AOMIC-PIOP2	226	226	896	0	896	628/90/178	Y	Y	Y
Beijing Enhanced	180	180	180	180	0	126/18/36	Y	Y	
Beijing Eyes Open Eyes Closed	48	48	144	0	144	99/15/30	Y	Y	Y
Beijing Short TR	27	27	109	0	109	76/12/21	Y	Y	
BHRC	578	578	825	0	825	578/82/165	Y	Y	
Caltech Conte Center	109	118	370	271	99	262/36/72	Y	Y	Y
CCNP	195	195	389	389	0	273/38/78	Y	Y	
CHCP (external)	171	171	2064	536	1528	1470/182/412			
CoRR	1357	1357	3477	3477	0	2432/349/696	Y	Y	
DLBS	452	452	1545	1545	0	1081/155/309	Y	Y	Y
ds003404	19	19	19	19	0	13/2/4	Y	Y	
FCP-1000	1023	1023	1023	1023	0	716/102/205	Y	Y	
GSP	1256	1256	2168	2168	0	1517/217/434	Y	Y	
HBN	1466	1466	2607	2607	0	1823/263/521	Y	Y	
HCP	625	625	10313	2278	8035	7164/982/2167	Y	Y	Y
ISYB	215	215	215	215	0	151/21/43	Y	Y	
LA5c (external)	265	265	1969	261	1708	1387/197/385			
NAD	298	298	595	595	0	416/60/119	Y	Y	
NKI	1312	1312	13036	6595	6411	9126/1304/2606	Y	Y	Y
PNC	1561	1561	4551	1495	3056	3189/452/910	Y	Y	Y
PPMI	558	808	808	808	0	562/80/166	Y	Y	Y
QTAB	413	413	2014	1417	597	1412/202/400	Y	Y	Y
QTIM	1123	1123	2290	1063	1227	1605/228/457	Y	Y	Y
SALD	493	493	493	493	0	345/49/99	Y	Y	
SLIM	587	587	1034	1034	0	725/103/206	Y	Y	
SRPBS	1406	1406	1406	1406	0	984/141/281	Y	Y	
Synaesthesia	127	127	507	507	0	355/52/100	Y	Y	
Total (registered)	21106	23691	63948	37121	26797				

is used for external testing. Longitudinal pairs link visits from the same subject, with a separate T1 scan for each visit. ADNI contributes 3,098 training pairs and PPMI contributes 204. The numbers of held-out pairs used for Phase-III evaluation are reported in Appendix C.3.

C Training and Evaluation Details

C.1 Phase-I

The structural and functional terms of the Phase-I loss (Section 3.4) are

$$\mathcal{L}_s = \lambda_s \, \|M \odot (\hat{x}^s - x^s)\|_1 + \lambda_{s,\mathrm{bg}} \, \|\bar{M} \odot \hat{x}^s\|_1 + \lambda_{\mathrm{ssim}} \big(1 - \mathrm{SSIM}(\hat{x}^s, x^s)\big) + \lambda_{s,\mathrm{wav}} \, \mathcal{L}_{\mathrm{wav}}(\hat{x}^s, x^s), \tag{10}$$

and

$$\begin{aligned} \mathcal{L}_f = {} & \lambda_f \, \|M \odot (\hat{x}^f - x^f)\|_1 + \lambda_{f,\mathrm{bg}} \, \|\bar{M} \odot \hat{x}^f\|_1 + \lambda_{f,\mathrm{wav}} \, \mathcal{L}_{\mathrm{wav}}(\hat{x}^f, x^f) \\ & + \lambda_\rho \big(1 - \rho_t(\hat{x}^f, x^f)\big) + \lambda_{\mathrm{dyn}} \, \mathcal{L}_{\mathrm{dyn}} + \lambda_\Delta \, \mathcal{L}_\Delta + \lambda_{\mathrm{var}} \, \mathcal{L}_{\mathrm{var}}, \end{aligned} \tag{11}$$

where M is the brain mask, $\bar{M}$ the region outside the brain, and $\mathcal{L}_{\text{wav}}$ the wavelet loss of Yasinzai et al. (2026). The background terms supervise the raw, unmasked decoder output, encouraging zero intensity outside the brain. The four temporal terms in $\mathcal{L}_f$ capture different aspects of the functional signal. The per-voxel temporal Pearson correlation ρ_t measures agreement between reconstructed and target time courses. $\mathcal{L}_{\text{dyn}}$ measures reconstruction error after temporal de-meaning and normalization by the target's dynamic RMS, emphasizing fluctuations rather than the static baseline. $\mathcal{L}_{\Delta}$ applies the same error to frame-to-frame differences. $\mathcal{L}_{\text{var}}$ is the squared log-ratio of reconstructed to target temporal variance and discourages over-smoothing.

The KL weight remains 10^{-4} throughout training. With this small weight, the latent is effectively deterministic, and subsequent phases use μ_t alone. We trained for 50 epochs with AdamW, sampling one random 32-frame window per run in each epoch. All later phases use the frozen checkpoint from step 40,000.

C.2 PHASE-II

- **Model.** We used the configuration described in Sections 3.2, 3.3, and Appendix A. The bridge connects the full previous window W_k to the next window W_{k+1}, with a single s shared across frames. We set the noise scale to $\epsilon = 0.206$, the root-mean-square latent increment over one window measured on the training split, rather than tuning it as a hyperparameter. This matches the noise scale to the variation observed over that interval. Too little noise risks making the mapping effectively deterministic, whereas too much noise can move intermediate states away from the data manifold.

- **Losses.** Spatial generation learns a conditional distribution, aiming to preserve variance, power spectra, and functional connectivity without requiring frame-by-frame agreement. Temporal generation directly optimizes temporal correlation; as a conditional-mean prediction, it has reduced variance. Both use the same velocity field, whose value at $s = 0$ gives the conditional-mean increment.

  We fixed half of each batch at $s = 0$ for the single-step prediction $\hat{W}_{k+1} = W_k + v_\theta(W_k, 0, c)$. For the other half, we sampled $s \sim U(0, 1]$ to learn the full transport field, which is integrated with 50 Euler–Maruyama steps at inference. The losses follow this division. $\mathcal{L}_{\text{corr}}$, the energy-weighted per-location temporal Pearson correlation, applies only at $s = 0$. It supervises the deterministic prediction without access to target information: for $s > 0$, the input I_s already contains W_{k+1}, allowing the model to improve correlation by using target information in the input. The window-mean loss $\mathcal{L}_{\text{mean}} = \|\overline{\hat{W}_{k+1}} - \overline{W_{k+1}}\|^2$ is also applied only at $s = 0$ and is omitted from the main text for brevity. It constrains the temporal mean, which correlation alone does not capture. $\mathcal{L}_{\text{stat}}$ combines temporal-spectrum and spatial-covariance losses. We apply it only at $s \geq 0.5$, where the estimate of W_{k+1} from I_s and v_θ is reliable. $\mathcal{L}_{\text{BAM}}$ is given in Appendix A.5.

- **Condition dropout.** Text, stimulus tracks, and individual prior pathways are dropped independently at the rates given in Appendix A.6. This trains the same model for conditional and unconditional generation and encourages the fast action to represent changes whose drivers are not provided as conditions.

- **Optimization.** We trained for 50 epochs on 6 H100 GPUs with an effective batch size of 48, using AdamW and cosine learning-rate decay. Downstream evaluation and post-training use the checkpoint at step 180,000.

- **Evaluation.** We selected one run per (subject, task) from the test split. Given an initial 32-frame window, the model generates 128 frames (four windows), using its own output as context for each subsequent window. All evaluation horizons are prefixes of the same rollout. Results at $h96$ and MSE are reported in Appendix E.2. For task runs, we evaluated the same checkpoint with *true* and *zero* stimulus inputs. Temporal generation is evaluated using ROI temporal correlation R (Schaefer-400 (Schaefer et al., 2018); averaged over ROIs, then runs, then subjects) and MSE. Resting runs use spatial generation and are evaluated with FC cosine similarity, log-PSD MAE, dFC cosine similarity, and the dFC variance ratio (Table 20). The baseline is the MONAI model described in Appendix D.2.

C.3 PHASE-III

- **Shared settings.** Phase III starts from the Phase-II checkpoint with a lower learning rate and otherwise uses the same optimization settings. One in five samples in each batch is drawn from Phase II to reduce forgetting. The slow posterior and prior are trained in this phase, following the alignment, mixing, and annealing procedures used for the fast action.
- **Rest-to-task.** We used HCP within-session pairs for internal experiments and CHCP for external evaluation, with temporal generation only. Training focuses on the first window, which contains the state transition. Later windows continue the same state, a task already learned in Phase II. We do not supply a stimulus track: HCP uses fixed block designs, and timing is conveyed through the task label and slow action. The identity amplitude term in Section 3.4 compares the per-voxel temporal standard deviation of the generated window with that of the real one, each centred on its own batch mean:

$$\mathcal{L}_{\mathrm{id}} = \mathrm{SmoothL1}\big(A(\hat{W}) - \overline{A(\hat{W})},\ A(W) - \overline{A(W)}\big), \qquad A(W)_v = \mathrm{std}_t\, W_{t,v}, \quad (12)$$

  where $\overline{\cdot}$ is the mean over the batch and v indexes locations. Batch centring removes the shared task pattern that dominates the flow loss, emphasizing subject-specific deviations that would otherwise receive little supervision. We trained for 8k steps and evaluated R between the generated and real first 32-frame windows under temporal generation. The task description is either *matched* (the task actually performed) or *mismatched* (another task of the same subject, from the same resting source).
- **Longitudinal.** We trained for 30 epochs using spatial generation on ADNI and PPMI longitudinal pairs with intervals $\geq$ 12 months, together with 20% resting-state replay from ABIDE-I. Change between visits is represented by the difference between their time-averaged latents, $\Delta z = \overline{z}_B - \overline{z}_A$. The structural change loss compares the generated and observed changes in both spatial pattern and magnitude:

$$\mathcal{L}_{\mathrm{struct}} = 1 - \rho\big(\Delta\hat{z},\, \Delta z\big) + \tfrac{1}{2}\Big(\tfrac{\|\Delta\hat{z}\|}{\|\Delta z\|} - 1\Big)^2. \quad (13)$$

  Spatial correlation ρ measures agreement in the pattern of change independently of its magnitude, which matters because anatomical change accounts for only 1–2% of the signal. The norm-ratio term separately constrains the magnitude. We computed this loss only on the first window. Continuation windows already contain the target visit's anatomy in their inputs, so they do not require the model to infer the anatomical change.
- **Evaluation.** We generated a follow-up scan from the source scan using spatial generation with three specified intervals: the observed interval, zero interval ($0\times$), and twice the observed interval ($2\times$). The generated latent is decoded into a T1 image for two analyses. For qualitative comparisons across intervals, we examine changes in each structure's intensity relative to the same scan's cortex. For quantitative evaluation, we segmented each structure and correlated its generated volume change with the observed change across subjects (Appendix E.3).

  Both analyses use the combined validation and test pairs from the two cohorts: 945 pairs from 269 subjects, comprising 859 ADNI pairs and 86 PPMI pairs. The test split alone contains too few longitudinal pairs to estimate per-structure correlations. No model selection is performed on the validation set in this phase: the 30-epoch training budget is fixed in advance, as is the 8k-step budget for rest-to-task training.

C.4 DOWNSTREAM TASKS

- **Prediction targets.** We evaluated 52 targets (28 primary and 24 secondary), covering demographics (age, sex), diagnosis (ASD, ADHD, AD, MCI, MDD, PD, BD, SCZ, four-class), cognition (Raven, five NEO scales), and state (rest vs. all tasks pooled as primary; rest vs. each task as secondary). We used the same subject-level 7:1:2 splits as in pre-training, with CHCP and LA5c as external cohorts. Stable phenotypes are predicted from resting runs only, with features averaged first across windows within each run and then across runs within each subject. State prediction is evaluated per run.

- **BrainWM features.** For the VAE, we took the brain-masked temporal mean of μ, applied adaptive pooling, and flattened the result. For the DiT, we zeroed the text and stimulus inputs and concatenated mean-pooled tokens from four intermediate layers at $s = 0$. DiT+BAM uses the same procedure with the posterior action token injected.
- **Strict linear probe.** We froze the features and fitted standardization on the training set. Classification uses L2-regularized logistic regression, and regression uses ridge regression. Regularization strength is selected using the primary validation metric. The probe is then refit on the combined training and validation sets and evaluated once on the test set. We obtained 95% confidence intervals from 2,000 subject-level bootstrap samples.
- **Full fine-tuning.** BrainWM and BrainLM use the same fixed protocol: one seed, no hyperparameter search, and 50 epochs without early stopping. We retained the best validation checkpoint from epochs 1–30 and from epochs 1–50, evaluating each once on the test set. The main text reports the latter. Both BrainLM results are listed as BrainLM-30 and BrainLM-50 in Appendix E.1. Classification uses balanced class weights calculated on the training set, with thresholds selected on the validation set.

D Baseline Models

D.1 Downstream Baselines

We evaluated five fMRI foundation models using their released checkpoints, preprocessing procedures, and atlases. All models use the same data splits, prediction targets, and strict linear-probe protocol. For full fine-tuning or model-specific adaptation, we followed each model's released protocol; training budgets therefore differ across model families. No baseline receives labels, metadata, or text as conditioning inputs.

Table 10: Downstream baselines and their adaptation protocols.

Model	Input	Pretraining objective	Checkpoint	Adaptation protocol
BrainLM (Caro et al., 2024)	ROI time series (A424)	masked autoencoding	released	fixed protocol: seed 42, lr 10^{-5}, 50 epochs, best within 30
Brain-JEPA (Dong et al., 2024)	ROI time series	JEPA + gradient positioning	UKB ViT-B, epoch 300	official end-to-end, seed 0, 50 epochs, lr $10^{-3} \times b/256$, layer-wise decay 0.75
BrainMass (Yang et al., 2024)	ROI functional connectivity	masked-ROI modeling + representation alignment	released	native: frozen BrainTF + RBF-SVM (C=1, gamma=scale)
NeuroSTORM (Wang et al., 2026a)	4D voxels	spatiotemporal self-supervision (28.65M frames)	released	native: task prompt tuning (50 prompts); full fine-tuning as a resource-matched secondary result
Brain-DiT (Xia et al., 2026a)	ROI	metadata-conditioned diffusion pretraining	released	native: frozen backbone + query aggregator; full fine-tuning as the official multi-scale end-to-end

D.2 Generative Baselines

MONAI-based conditional 4D latent diffusion. This baseline combines a framewise 3D VQ-VAE from MONAI (Pinaya et al., 2023) with a conditional DiT3D. We trained the VQ-VAE from scratch on the same Phase-I training set. The DiT3D is conditioned on three inputs: a fixed source memory obtained by query-resampling the initial 32-frame context, the previous 32-frame latent window, and a frozen 768-dimensional text embedding supplied through cross-attention in every

block. We used DDPM training with 1000 diffusion steps. Evaluation follows the same rollout protocol as BrainWM: observe frames $[0, 32)$, then generate subsequent windows autoregressively.

Baselines without additional training. We also evaluated four baselines that do not use a learned generator. Their outputs follow the same decoding, masking, parcellation, and scoring pipeline as BrainWM. The first three construct future latent windows using fixed rules. Let W_0 denote the observed 32-frame window, m its temporal mean, and $H = 128$ the forecasting horizon.

- **repeat** repeats W_0 to fill the horizon, using four copies for $H = 128$. This tests whether generation improves on simply replaying the observed activity. It is a particularly important reference for connectivity metrics because it retains the observed window's interregional covariance and variance, while imposing a repeating temporal pattern.
- **window mean** outputs m at every frame. As a constant prediction, it tests how well a metric can be satisfied without any temporal variation. Its temporal variance is zero, so temporal correlation and both connectivity measures are undefined, not zero. Spectral and variance metrics can still be evaluated against this constant-signal reference.
- **AR(1)** estimates a lag-one autocorrelation a for each latent unit from W_0 and clips it to $[-0.99, 0.99]$. It extrapolates from the last observed frame towards m using $x_t = m + a^t(x_{\text{last}} - m)$. This provides a standard reference for autocorrelated signals that decay towards their mean rather than repeat a fixed window. It also illustrates a limitation of connectivity similarity: similar decaying trajectories across regions can produce correlations near ± 1 without reproducing realistic dynamics (Appendix E.2).
- **no simulation** is a reference for longitudinal simulation, not forecasting. It reconstructs the source scan with the frozen autoencoder without simulating a change. We scored this reconstruction with the same metrics as the generated follow-up, providing a no-change reference for the reported change correlations.

E Additional Results

E.1 Representing the Brains

We supplement Table 1 with results on the remaining three downstream tasks under full fine-tuning (Table 11) and on all eight tasks under strict linear probing (Table 12). We followed the settings in Section 4.2. Baselines use their official checkpoints, and all three BrainWM feature variants are extracted with text and stimulus conditions set to zero. BrainWM and BrainLM follow the same fixed fine-tuning protocol (Appendix C.4). Linear probes use L2-regularized logistic regression or ridge regression on frozen features. Bold values indicate the best result overall; underlined values indicate the best baseline result.

Table 11: Full fine-tuning results on the three downstream tasks not included in the main table. Bold indicates the best result overall; underlining indicates the best baseline result.

	ABIDE-I	AOMIC	LA5c (external)
	Sex	NEO-N	HC vs ADHD
Model	AUC↑ / F1↑	RMSE↓ / r ↑	AUC↑ / F1↑
BrainLM-30	0.572 / 0.473	<u>6.706</u> / 0.166	0.438 / 0.275
BrainLM-50	0.572 / 0.473	6.722 / 0.166	0.453 / 0.401
Brain-JEPA	0.585 / 0.398	6.798 / 0.061	0.354 / 0.327
BrainMass	0.583 / 0.456	6.781 / 0.071	**0.714** / 0.429
NeuroSTORM	0.598 / 0.410	6.797 / <u>0.178</u>	0.430 / 0.442
Brain-DiT	<u>0.606</u> / <u>0.571</u>	6.855 / 0.044	0.526 / <u>0.588</u>
BrainWM-VAE	**0.687** / 0.538	**6.246** / **0.438**	0.688 / **0.610**
BrainWM-DiT	0.615 / 0.547	6.870 / 0.229	0.448 / 0.343
BrainWM-DiT+BAM	0.591 / **0.584**	7.437 / 0.188	0.547 / 0.491

Table 12: Strict linear-probe results on all eight downstream tasks using frozen features, in two blocks of four. Bold indicates the best result overall; underlining indicates the best baseline result.

	ABIDE-I		AOMIC	ADNI
Model	Age RMSE↓ / r ↑	Sex AUC↑ / F1↑	NEO-N RMSE↓ / r ↑	HC vs AD AUC↑ / F1↑
BrainLM	5.581 / 0.635	0.544 / 0.403	6.761 / 0.130	0.631 / 0.579
Brain-JEPA	5.105 / 0.710	0.580 / 0.518	6.918 / −0.053	0.567 / 0.585
BrainMass	5.616 / 0.616	0.593 / 0.480	6.848 / 0.033	0.622 / 0.585
NeuroSTORM	4.620 / 0.772	0.626 / **0.539**	6.609 / **0.516**	0.650 / 0.564
Brain-DiT	4.719 / 0.749	0.686 / 0.538	7.550 / 0.053	0.667 / 0.609
BrainWM-VAE	**3.647** / **0.859**	**0.687** / 0.498	6.684 / 0.352	**0.828** / **0.639**
BrainWM-DiT	4.135 / 0.831	0.618 / 0.477	6.721 / 0.365	0.715 / 0.619
BrainWM-DiT+BAM	4.071 / 0.836	0.609 / 0.534	**6.380** / 0.398	0.733 / 0.586

	SRPBS		LA5c (external)	
Model	HC vs MDD AUC↑ / F1↑	HC vs SCZ AUC↑ / F1↑	HC vs BD AUC↑ / F1↑	HC vs ADHD AUC↑ / F1↑
BrainLM	0.716 / 0.623	0.771 / 0.637	0.575 / 0.575	0.646 / 0.508
Brain-JEPA	**0.748** / **0.662**	0.756 / 0.605	0.429 / 0.393	0.380 / 0.263
BrainMass	0.675 / 0.584	0.734 / 0.553	0.567 / 0.575	0.401 / 0.515
NeuroSTORM	0.696 / 0.605	0.741 / 0.601	**0.658** / 0.414	0.464 / 0.401
Brain-DiT	0.731 / 0.647	0.731 / 0.610	0.654 / **0.652**	0.505 / 0.459
BrainWM-VAE	0.694 / 0.623	0.827 / 0.703	0.346 / 0.490	**0.703** / 0.435
BrainWM-DiT	0.656 / 0.594	**0.849** / **0.731**	0.613 / 0.589	0.526 / 0.338
BrainWM-DiT+BAM	0.678 / 0.599	0.835 / 0.707	0.621 / 0.572	0.479 / **0.559**

Results. With fine-tuning, BrainWM-VAE achieves the highest sex classification AUC and the best NEO neuroticism regression results among the three remaining tasks. For neuroticism, r increases from 0.18 for the best baseline to 0.44. ADHD classification is the only one of the eight tasks in which a baseline has the highest AUC: BrainMass reached 0.714, compared with 0.688 for BrainWM-VAE. BrainWM-VAE nevertheless achieved the highest F1 on this task.

Under linear probing, a BrainWM feature variant led on six of the eight tasks, with AUCs of 0.83 for AD and 0.85 for SCZ. Brain-JEPA performs best on MDD (0.748 vs. 0.694), and NeuroSTORM on BD (0.658 vs. 0.621). The relative strengths of the BrainWM features are broadly consistent across linear probing and fine-tuning: VAE features perform best on anatomy-related targets such as age and AD, whereas DiT and DiT+BAM features perform better on SCZ and BD.

E.2 Forecasting the Brains

Results. Tables 18 to 20 provide the dataset-, task-, and horizon-level results underlying the two forecasting summary tables in the main text. Tables 13 to 15 compare BrainWM with the three fixed-rule predictors described in Appendix D.2.

We evaluated 19 dataset–task settings at four horizons, giving 76 comparisons with each baseline. BrainWM achieved higher temporal correlation than repeat in 74 of 76 comparisons and than AR(1) in 65 of 76. At the shortest horizon, $h32$, it outperforms both in all 19 settings. Across the 328 subjects with task runs, its temporal-correlation advantage is $+0.139$ over repeat and $+0.081$ over AR(1) at $h32$, decreasing to $+0.055$ and $+0.037$ at $h128$. All four differences have $p < 0.0001$.

For resting runs, we focused on generative metrics rather than temporal correlation, since there is no task paradigm whose time course the model can follow (Table 15). BrainWM has higher FC cosine similarity than repeat in all four cohorts, with gains of $+0.068$, $+0.016$, $+0.080$, and $+0.052$, and substantially lower spectral error. For dynamic connectivity, BrainWM achieved cosine similarities of 0.61–0.65, compared with 0.16–0.30 for AR(1).

We also evaluated the four 32-frame blocks of each rollout separately, comparing each with the corresponding real frames rather than evaluating cumulative prefixes (Table 16). Most of the advantage over repeat is concentrated in the first block. The advantage over AR(1) remained significant in all four blocks.

All statistical analyses used subjects as the unit of analysis. We first average each subject's observations, then calculate confidence intervals from 2,000 subject-level bootstrap samples. Paired comparisons use 10,000 sign-flip permutations of the per-subject differences. We distinguish between two aggregation schemes: setting-level summaries weight the 19 dataset–task settings equally, whereas pooled summaries weight the 779 subjects equally. These schemes can produce substantially different values; for example, the same advantage is $+0.143$ under setting-level averaging and $+0.081$ under subject-level averaging. We kept the aggregation scheme consistent within each comparison.

Limitations. BrainWM performs less well on two measures of second-order signal statistics in Table 14. Its lag-one autocorrelation ratio is 0.7271, further from the reference value of one than repeat's 0.7704; the paired difference of -0.0433 is statistically significant. Its variance retention is 2.1891, indicating more than twice the temporal variance of the real signal. Repeat inherits these statistics from an observed window, but the comparison still identified limitations in the generated dynamics.

The block-level results are consistent with error accumulation during rollout. In Table 17, variance retention increases with each successive block under both generation modes, whereas repeat stayed near 1.05. Only the first generated block is conditioned on real observations; subsequent blocks use generated inputs. This also helps explain why the advantage over repeat is largest in the first block.

The settings where AR(1) outperforms BrainWM are concentrated at longer horizons: CHCP emotion and language from $h64$, HCP emotion from $h96$, and resting runs at the longest horizons. Without a task paradigm to follow, a predictor that decays smoothly towards the observed mean can compare favorably with a rollout that accumulates error.

Comparisons also require attention to undefined metrics. The constant window-mean prediction has undefined temporal correlation and functional connectivity, not zero values. AR(1) yields valid connectivity estimates for only 179 subjects, so its aggregate FC score should not be compared with a BrainWM score calculated over a different set of subjects. On the paired subset, its advantage is not statistically significant.

Finally, these results do not establish subject-specific forecasting. Using a generated run as a query and real runs of the same task from a different session as the gallery, FC identified the correct subject 17.0% of the time, above the 3.0% chance rate. However, identification accuracy is 50.1% using the 32 observed input frames and 80.0% using a real continuation. The rollout therefore retained only part of the individual information available in its input. Time-course-based identification reached twice chance but also remained below the input baseline.

E.3 Simulating the Brains

E.3.1 Rest-to-task

Results. We evaluated rest-to-task simulation in terms of temporal agreement and spatial patterns, using reference measurements to interpret each result.

For temporal agreement, Table 21 compares the generated window with three references: the leave-one-subject-out mean trajectory for the same task, a second real run from the same subject performing that task, and a real run from another subject. All are evaluated on the same pairs and the same first 32-frame window using temporal correlation.

For spatial patterns, we construct a 400-region map of task-window fluctuation amplitudes, expressed as deviations from the task's typical pattern. The generated map correlates with the subject's real map at 0.227. Replacing the source with another subject's resting scan yields a negative correlation, giving a paired difference of 0.275 ($p < 0.001$, $n = 121$). In subject identification among 121 candidates, the generated map achieved 67.9% accuracy, compared with 1.7% chance and 96.0% when a second real scan is used as the query. To avoid shared estimation error, we estimate the

Table 13: Comparison with fixed-rule predictors, averaged equally across settings. Temporal correlation is reported for the 15 task settings used in Section 4.3 and for all 19 settings. Signal statistics are reported for all 19 settings, including task and resting runs; resting-only results are given in Table 15. The reference value for variance retention is 1.0. Dashes indicate metrics undefined for a constant series. Additional metrics and paired tests are reported in Table 14.

	Temporal correlation $R \uparrow$				Signal statistics		
Method	$h32$	$h64$	$h96$	$h128$	FC cos↑	PSD MAE↓	var. ret.→ 1
15 task settings							
BrainWM	**0.166**	**0.107**	**0.086**	**0.070**			
repeat	−0.043	−0.012	−0.004	−0.007			
AR(1)	0.043	0.037	0.033	0.030			
all 19 settings							
BrainWM	**0.143**	**0.093**	**0.073**	**0.060**	0.692	**2.006**	2.281
repeat	−0.035	−0.009	−0.002	−0.005	0.639	15.547	**1.062**
AR(1)	0.040	0.034	0.030	0.027	**0.854**	7.090	0.002
window mean	—	—	—	—	—	12.903	0.000

Table 14: Subject-level comparison with fixed-rule predictors, pooled across $n = 779$ subjects for temporal correlation. The BrainWM column reports absolute scores; the next two columns report paired differences (BrainWM − baseline), tested using sign-flip permutations. Negative log-PSD MAE differences favor BrainWM. For paired comparisons, n counts subjects with valid estimates for both methods. FC and dFC have smaller samples because these metrics are undefined for degenerate time series. $^{***}p < 0.0001$.

	BrainWM	− repeat	− AR(1)	n
Temporal correlation				
$h32$	0.0805	$+0.0747^{***}$	$+0.0495^{***}$	779
$h64$	0.0549	$+0.0448^{***}$	$+0.0352^{***}$	779
$h96$	0.0455	$+0.0362^{***}$	$+0.0301^{***}$	779
$h128$	0.0423	$+0.0361^{***}$	$+0.0296^{***}$	779
Signal statistics of the generated run				
FC cosine	0.7254	$+0.0400^{***}$	−0.0097 (n.s., $p = 0.73$)	179
dFC variability ratio	0.6493	$+0.2548^{***}$	-8.2869^{***}	147
variance retention	2.1891	$+1.4103^{***}$	$+2.1871^{***}$	779
lag-1 autocorr. ratio	0.7271	-0.0433^{***}	-0.1060^{***}	779
log-PSD MAE	1.8941	-11.5748^{***}	-5.2416^{***}	779

reference task means on separate data halves: one is subtracted from the generated maps and the other from the real maps.

Limitations. The group task average has the highest temporal correlation in Table 21, exceeding the subject-specific references. This suggests that, over a single 23-second window, the shared task response accounts for much of the predictable time course. We therefore do not interpret temporal agreement as evidence of subject-specific prediction.

The spatial analysis must also account for information already present in the source resting scan. Scoring that scan directly under the same centring procedure gave a correlation of 0.2414, compared with 0.2267 for the generated window. The difference of -0.0146 is not significant. The result thus supports preservation of an individual spatial pattern across the state transition, rather than the generation of additional individual information.

Two further limitations apply. Amplitude maps from two generated samples of the same pair correlate at 0.989, compared with 0.421 for two real scans from the same subject. The generated mapping is therefore nearly deterministic given its source; this analysis does not demonstrate realistic sample diversity. In addition, the spatial analysis has been conducted only on HCP, not on the external cohort.

Table 15: Resting-state comparisons by cohort, using the evaluation pipeline of Table 3. Variance retention, dFC variability ratio, and lag-one autocorrelation ratio have a reference value of 1.0. Dashes indicate unavailable estimates. FC and dFC require at least two regions with variance above the estimator's threshold. AR(1) decayed below this threshold in most runs, leaving estimates for 49 of 198 ABIDE-I subjects, 41 of 281 SRPBS subjects, 14 of 123 HCP subjects, and none of the 32 CHCP subjects. The constant window-mean prediction has no valid FC or dFC estimate.

Metric	Method	ABIDE-I	SRPBS	HCP rest	CHCP rest
FC cosine ↑	BrainWM	0.746	0.755	0.676	0.692
	repeat	0.677	0.739	0.596	0.640
	AR(1)	0.652	0.696	0.753	—
log-PSD MAE ↓	BrainWM	1.838	1.944	1.892	2.045
	repeat	10.94	12.54	16.23	15.86
	AR(1)	7.216	6.809	7.369	7.221
	window mean	13.30	12.50	—	—
dFC cosine ↑	BrainWM	0.609	0.654	0.618	0.647
	repeat	0.586	0.674	0.569	0.614
	AR(1)	0.199	0.157	0.297	—
dFC var. ratio → 1	BrainWM	0.579	0.697	0.732	0.642
	repeat	0.352	0.428	0.396	0.340
	AR(1)	2.302	3.049	159.9	—
variance retention → 1	BrainWM	1.792	2.667	1.509	2.925
	repeat	0.719	0.696	1.049	1.479
	AR(1)	0.002	0.002	0.001	0.001
	window mean	1.511	2.869	0.000	0.000
lag-1 ratio → 1	BrainWM	0.772	0.731	0.649	0.655
	repeat	0.803	0.754	0.756	0.733
	AR(1)	0.982	0.787	0.684	0.586

Table 16: Temporal correlation evaluated separately on each 32-frame block, rather than on the cumulative prefixes used in Table 2. Differences are paired over 779 subjects. $^{**}p < 0.01$, $^{***}p < 0.0001$.

Frames	BrainWM	repeat	− repeat	− AR(1)
1–32	0.0805	0.0058	$+0.0747^{***}$	$+0.0495^{***}$
33–64	0.0266	0.0172	+0.0094 (n.s.)	$+0.0233^{***}$
65–96	0.0183	0.0113	+0.0071 (n.s.)	$+0.0402^{***}$
97–128	0.0241	0.0017	$+0.0223^{**}$	$+0.0351^{***}$

Table 17: Generated-to-real temporal variance ratio for each rollout block, reported as the median across ROIs. Ratios computed from total variance across regions show the same increasing trend: 5.0/14.0/24.7/35.3 for temporal generation and 2.33/3.58/4.99/6.36 for spatial generation.

Frames	temporal generation	spatial generation	repeat
1–32	5.16	2.66	1.09
33–64	12.94	4.25	1.09
65–96	21.39	6.05	1.07
97–128	27.84	7.78	1.04

E.3.2 LONGITUDINAL

Results. We measured structural volumes on 945 held-out pairs from 269 subjects: 859 pairs from ADNI and 86 from PPMI. All scans are segmented with SynthSeg. The held-out set combines validation and test pairs for the reason given in Appendix C.3.

Table 18: Forecasting results on internal cohorts: all seven HCP tasks and ABIDE-I rest. R denotes ROI temporal correlation over each cumulative horizon; MSE is reported at $h128$. BrainWM true and zero use the actual stimulus track and a zeroed track, respectively. Bold indicates the best value in each column within a task.

Task	Model	$R\ h32$	$R\ h64$	$R\ h96$	$R\ h128$	MSE $h128$
HCP emotion	BrainWM true	0.090	0.028	0.021	0.011	**0.0105**
	BrainWM zero	**0.093**	**0.033**	**0.024**	**0.014**	0.0106
	MONAI	0.001	0.005	0.005	0.008	0.0174
HCP gambling	BrainWM true	**0.169**	**0.119**	**0.093**	**0.070**	**0.0150**
	BrainWM zero	0.162	0.111	0.089	0.068	0.0151
	MONAI	0.002	0.002	0.004	0.005	0.0169
HCP language	BrainWM true	**0.181**	**0.089**	**0.072**	**0.051**	0.0120
	BrainWM zero	0.175	0.073	0.056	0.037	**0.0117**
	MONAI	0.003	0.013	0.012	0.016	0.0172
HCP motor	BrainWM true	**0.163**	**0.090**	**0.059**	**0.063**	**0.0130**
	BrainWM zero	0.158	0.082	0.044	0.043	0.0132
	MONAI	0.002	0.002	0.009	0.019	0.0173
HCP relational	BrainWM true	**0.146**	**0.179**	**0.140**	**0.111**	**0.0132**
	BrainWM zero	0.138	0.170	0.130	0.101	0.0134
	MONAI	0.001	0.009	0.008	0.014	0.0169
HCP social	BrainWM true	0.348	**0.241**	**0.197**	**0.155**	**0.0151**
	BrainWM zero	**0.351**	0.240	0.196	0.150	0.0156
	MONAI	0.005	−0.001	0.012	0.013	0.0169
HCP working memory	BrainWM true	**0.175**	**0.093**	**0.089**	**0.097**	**0.0103**
	BrainWM zero	0.170	0.083	0.079	0.087	**0.0103**
	MONAI	0.000	0.009	0.007	0.005	0.0167
ABIDE-I rest	BrainWM zero	**0.068**	**0.042**	**0.035**	**0.041**	**0.0097**
	MONAI	0.000	0.002	0.003	0.008	0.0182

The analysis was specified before examining individual structures. For each structure, we correlate the generated source-to-follow-up volume change with the observed change across subjects. Before calculating this correlation, we adjust both changes for three variables: follow-up interval, cohort, and the structure's baseline volume. Interval adjustment accounts for the greater change expected over longer follow-up periods. Cohort adjustment prevents differences between ADNI and PPMI, such as their several-fold difference in mean hippocampal change, from driving the correlation. Baseline-volume adjustment accounts for the tendency of larger structures to show larger absolute changes, information already available in the model's input. We first aggregate multiple follow-up visits within each subject and obtain confidence intervals from $4{,}000$ subject-level bootstrap samples.

Two structures have confidence intervals that exclude zero (Table 22 and Figure 6): the temporal horn of the lateral ventricle (0.187) and the hippocampus (0.137). These neighboring structures reflect medial temporal atrophy and expansion of the adjacent cerebrospinal fluid space.

Limitations. To assess how much change can be measured through the reconstruction pipeline, we computed the same correlation using a real follow-up passed through the frozen autoencoder as a reconstruction reference. This correlation is only 0.32 for the amygdala, 0.23 for the thalamus, and 0.23 for cerebral white matter, indicating that much of the observed change is not preserved by the measurement pipeline. Low generated correlations for these structures are therefore difficult to attribute to the world model alone. The cerebral cortex is different: its reconstruction reference is 0.52, but the generated correlation is -0.08, showing that the model fails to capture cortical volume change despite its measurability through the pipeline.

The ventricular results are sensitive to baseline-volume adjustment. For total ventricular volume, summed over all four ventricles, the correlation is 0.19 $[0.08, 0.30]$ after adjusting for interval and

Table 19: Forecasting results on the external CHCP cohort, covering all seven tasks and rest. Metrics and notation follow Table 18.

Task	Model	$R\,h32$	$R\,h64$	$R\,h96$	$R\,h128$	MSE $h128$
CHCP emotion	BrainWM true	**0.101**	0.023	−0.003	−0.004	**0.0086**
	BrainWM zero	0.097	**0.024**	**0.010**	**0.015**	**0.0086**
	MONAI	−0.003	0.005	0.006	0.009	0.0179
CHCP gambling	BrainWM true	0.068	0.115	0.089	0.070	0.0253
	BrainWM zero	**0.118**	**0.145**	**0.128**	**0.100**	**0.0127**
	MONAI	0.004	0.006	0.009	0.010	0.0175
CHCP language	BrainWM true	0.106	0.032	**0.036**	**0.025**	0.0575
	BrainWM zero	**0.121**	**0.042**	0.011	0.015	**0.0090**
	MONAI	0.002	0.010	0.016	0.018	0.0187
CHCP motor	BrainWM true	**0.149**	**0.067**	**0.083**	**0.065**	**0.0097**
	BrainWM zero	0.140	0.040	0.052	0.028	0.0107
	MONAI	0.008	0.005	0.011	0.024	0.0174
CHCP relational	BrainWM true	**0.193**	0.216	**0.149**	**0.103**	**0.0105**
	BrainWM zero	0.183	**0.217**	0.143	0.096	**0.0105**
	MONAI	0.003	0.005	0.006	0.018	0.0177
CHCP social	BrainWM true	**0.329**	**0.203**	**0.189**	**0.147**	**0.0120**
	BrainWM zero	**0.329**	0.198	0.185	0.143	0.0123
	MONAI	0.006	−0.005	0.005	0.005	0.0172
CHCP working memory	BrainWM true	**0.203**	**0.055**	0.040	0.040	0.0086
	BrainWM zero	0.193	0.036	**0.041**	**0.043**	**0.0085**
	MONAI	−0.003	0.013	0.008	0.006	0.0180
CHCP rest	BrainWM zero	**0.040**	**0.048**	**0.025**	**0.007**	**0.0089**
	MONAI	0.008	0.003	0.006	0.006	0.0176

Table 20: Resting-state generation over 128 frames, reported by cohort. Metrics are FC cosine similarity, log-PSD mean absolute error, dFC variability ratio, and dFC cosine similarity. CHCP is external. Bold indicates the better result for each comparison; a variance ratio closer to 1 is better.

Metric	Model	ABIDE-I	SRPBS	HCP rest	CHCP rest (ext.)
FC cosine ↑	BrainWM	**0.746**	**0.755**	**0.676**	**0.692**
	MONAI	0.180	0.170	0.148	0.196
log-PSD MAE ↓	BrainWM	**1.838**	**1.944**	**1.892**	**2.045**
	MONAI	2.623	3.229	2.538	2.817
dFC var. ratio → 1	BrainWM	0.579	0.697	**0.732**	**0.642**
	MONAI	**0.800**	**0.900**	0.646	0.573
dFC cosine ↑	BrainWM	**0.609**	**0.654**	**0.618**	**0.647**
	MONAI	0.067	0.068	0.112	0.140

cohort. It fell to 0.09 $[-0.03, 0.22]$ when baseline volume is also included. The lateral ventricle showed a similar decrease, from 0.14 to 0.04. These reductions suggest that much of the initial association reflects baseline size rather than predicted individual change. The hippocampal result is largely unchanged by this adjustment, from 0.143 to 0.137.

F Ablation Studies

We distinguish three types of intervention, each addressing a different question. *Input* interventions change the inputs to the released model, either by zeroing the stimulus track or by shifting acquisition times. *Pathway* interventions block an internal route by zeroing the latent action. *Model*

Table 21: Temporal correlation for rest-to-task generation and reference predictions, evaluated on the same pairs and first 32-frame windows. The group task average is the leave-one-subject-out mean trajectory for the same task. The other-subject reference uses a real scan from another subject performing that task.

	HCP	CHCP (external)
group task average	0.2411	0.2243
real test–retest	0.1342	0.1846
BrainWM, matched	0.154	0.131
another subject's scan	0.0864	0.0920

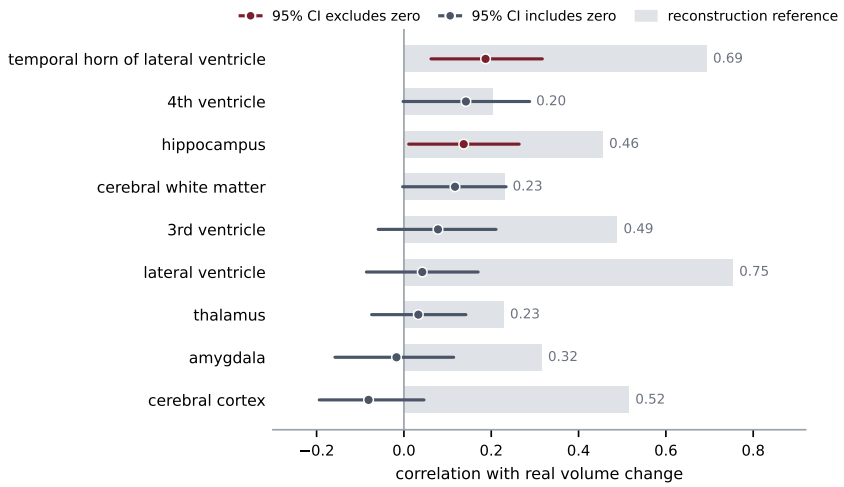

Figure 6: Longitudinal volume-change correlations by structure. Dots show correlations between generated and observed changes after adjustment for follow-up interval, cohort, and baseline volume. Error bars show 95% confidence intervals from 4,000 subject-level bootstrap samples. Grey bars show the reconstruction reference: the same correlation using a real follow-up passed through the frozen autoencoder. Low reference values indicate limited sensitivity of the reconstruction and measurement pipeline to change in that structure.

interventions remove a component and retrain from scratch with the same training budget. Retraining is needed to assess BAM's contribution during training: zeroing the action only at inference tests the trained model under a condition it has not encountered during training. Input and pathway interventions, in contrast, let us examine the model used for the main results without changing its weights.

Table 23 compares three trained models on the 14 task settings available for all three. The no-BAM configuration, `configs/phase2_nobam.yaml`, disables the latent action and supplies conditions to the DiT through a direct branch. It also removes the memory summary, which normally enters through the prior. This is therefore a no-BAM comparison, not a direct-stimulus-only comparison. With the real stimulus supplied, no BAM performs similarly to BrainWM. With the stimulus withheld, it fell behind by 0.031 to 0.039 across horizons, suggesting that BAM reduces dependence on an available event file.

The prior-only configuration, `configs/phase2_prioronly.yaml`, retains the eight-dimensional action bottleneck but does not train a posterior. It performs worse than both BrainWM and no BAM. These three models let us separate the benefit of posterior supervision (BrainWM minus prior only) from the effect of the bottleneck without that supervision (prior only minus no BAM). Posterior supervision improves performance by $+0.021$ to $+0.054$, with significant gains in all eight comparisons. With the real stimulus supplied, the bottleneck alone reduces performance by 0.021 to 0.041, significantly at three of the four horizons. Without the stimulus, its effect is not

Table 22: Correlations between generated and observed volume changes after adjustment for follow-up interval, cohort, and baseline volume. The reconstruction reference is obtained from a real follow-up passed through the frozen autoencoder and the same measurement pipeline. Low reference values limit the interpretation of generated correlations. All results use $n = 269$ subjects.

Structure	r	95% CI	Reconstruction reference
temporal horn of lateral ventricle	**0.187**	[0.06, 0.32]	0.693
4th ventricle	0.142	[−0.00, 0.29]	0.203
hippocampus	**0.137**	[0.01, 0.26]	0.455
cerebral white matter	0.117	[−0.00, 0.23]	0.230
3rd ventricle	0.078	[−0.06, 0.21]	0.488
lateral ventricle	0.042	[−0.09, 0.17]	0.753
thalamus	0.033	[−0.07, 0.14]	0.228
amygdala	−0.017	[−0.16, 0.11]	0.315
cerebral cortex	−0.081	[−0.19, 0.05]	0.515

significant. Thus, when a rich frame-level stimulus is available, the benefit of posterior supervision largely offsets the cost of compressing that signal through the action bottleneck. Spatial generation showed the same ordering: across the four resting settings, FC cosine similarity is 0.717 for BrainWM, 0.704 for no BAM, and 0.671 for prior only.

Table 24 instead tests the released model on seven HCP tasks, using 280 pairs per condition. Zeroing the action provides a check on the intervention: scores with the real and zero stimulus tracks were nearly identical at four-decimal precision. This is expected because the stimulus reaches the generator only through the action. When the real stimulus is supplied, zeroing the action reduces performance by 0.005 to 0.012, significantly at every horizon. The model therefore uses the action during inference.

Table 23: Temporal-generation results for three trained models, averaged equally over the 14 task settings shared by all models. Prior only retains BAM's action bottleneck but does not train a posterior. The lower block decomposes the BrainWM–no-BAM difference into the benefit of posterior supervision (BrainWM − prior only) and the effect of the bottleneck without that supervision (prior only − no BAM). Significance is assessed with paired t-tests over the 14 settings. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

	$h32$	$h64$	$h96$	$h128$
Stimulus given				
BrainWM	0.173	0.111	0.090	0.072
no BAM	0.169	0.117	0.090	0.077
prior only	0.128	0.082	0.069	0.049
Stimulus withheld				
BrainWM	0.173	0.107	0.085	0.067
no BAM	0.134	0.076	0.045	0.035
prior only	0.119	0.063	0.045	0.026
Stimulus given				
posterior supervision	$+0.045^{***}$	$+0.028^{***}$	$+0.021^{*}$	$+0.023^{**}$
the bottleneck	-0.041^{**}	-0.035^{**}	−0.021	-0.029^{*}
net (BrainWM − no BAM)	+0.004	−0.007	−0.001	−0.006
Stimulus withheld				
posterior supervision	$+0.054^{***}$	$+0.044^{***}$	$+0.040^{***}$	$+0.042^{***}$
the bottleneck	−0.015	−0.012	−0.001	−0.009
net (BrainWM − no BAM)	$+0.039^{***}$	$+0.031^{**}$	$+0.039^{***}$	$+0.032^{**}$

Table 24: Inference-time ablations on the released model. The noact condition sets the latent action to zero. The shift condition moves source and target acquisition times together by 48 frames (34.6 s), preserving their relative interval and leaving the stimulus track unchanged. It therefore disrupts alignment between absolute acquisition time and the block design while keeping timestamp values within the training range. Lower rows report differences from the corresponding base condition. $^{*}p < 0.05$, $^{***}p < 0.001$.

Condition	$h32$	$h64$	$h96$	$h128$
base, stimulus zero	0.1842	0.1208	0.0943	0.0801
noact, stimulus zero	0.1821	0.1166	0.0918	0.0758
shift, stimulus zero	0.1179	0.0443	0.0232	0.0216
base, stimulus true	0.1873	0.1255	0.1024	0.0875
noact, stimulus true	0.1822	0.1167	0.0917	0.0758
shift, stimulus true	0.1266	0.0574	0.0327	0.0269
base − noact, zero	+0.0021	+0.0042	+0.0025	+0.0044
base − noact, true	$+0.0051^{*}$	$+0.0088^{***}$	$+0.0107^{***}$	$+0.0117^{***}$
base − shift, zero	$+0.0662^{***}$	$+0.0765^{***}$	$+0.0711^{***}$	$+0.0586^{***}$
base − shift, true	$+0.0607^{***}$	$+0.0681^{***}$	$+0.0696^{***}$	$+0.0606^{***}$