ADWM: A LATENT WORLD MODEL FOR MODELING ALZHEIMER'S DISEASE PROGRESSION FROM FUNCTIONAL IMAGING

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding dynamic changes in brain function is essential for characterizing disease progression in Alzheimer's disease (AD). As a volumetric functional imaging modality, positron emission tomography (PET) imaging provides a direct view of cerebral metabolism and can capture disease-related metabolic alterations across brain regions. To model these evolving patterns, we introduce ADWM, a latent world model that learns the dynamics of dementia progression from sparse longitudinal observations. ADWM represents the PET modality and standardized uptake value ratio (SUVR) values in an aligned latent space, organized by a weakly supervised disease progression index (DPI). A shared neural ordinary differential equation (ODE) models population-level latent evolution along the DPI, while a separate nonnegative progression rate links disease stage to chronological time under longitudinal supervision. Conditioned on the initial state, ODE prediction, patient context, and visit-level treatment information, a residual diffusion model learns the distribution of individual latent changes beyond the shared progression trajectory. By combining a population-level progression prior with individual-level stochastic refinement, ADWM enables conditional simulation of latent disease states and regional metabolic profiles from sparse observations. This framework provides a functional-imaging-based world model for characterizing both shared and individualized AD progression dynamics, offering a unified approach to simulating disease evolution and its associated metabolic changes.

1 INTRODUCTION

Functional imaging, particularly positron emission tomography (PET), provides measurements of cerebral metabolism that capture disease-related changes across brain regions over the course of disease progression. In Alzheimer's disease (AD), longitudinal metabolic changes reflect the evolving functional state of the brain and provide information beyond a single-time-point assessment. These observations can be viewed as partial observations of an underlying disease state that continuously evolves. Modeling AD progression from functional imaging therefore requires not only predicting future observations, but also representing disease states and the dynamics governing their evolution.

A key challenge is that AD progression varies across individuals in both metabolic state and rate of evolution. Sparse and irregular follow-up observations further complicate the estimation of coherent progression trajectories Liu et al. (2025) Sharma et al. (2026). Chronological time does not necessarily correspond to a consistent disease stage across individuals, while observed metabolic changes also contain patient-specific variation. A functional-imaging-based world model should therefore capture a shared progression structure while allowing disease states to evolve at different rates and exhibit individual-specific deviations. This motivates decomposing disease-state evolution into shared progression dynamics and patient-specific stochastic refinement.

Recent generative medical models have explored conditional disease simulation and longitudinal image forecasting Yang et al. (2025); Puglisi et al. (2025), demonstrating the potential to synthesize plausible future observations from incomplete longitudinal data. However, directly modeling the full distribution of image changes may entangle shared disease progression with patient-specific variability, making the underlying transition dynamics difficult to characterize explicitly. A disease

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.

world model should instead establish a structured progression prior and model individual deviations around this shared trajectory.

To this end, we propose ADWM, a latent world model for modeling AD progression from sparse longitudinal functional imaging observations. We construct a joint PET–SUVR (standardized uptake value ratio) representation in an aligned latent space and organize its evolution with a weakly supervised disease progression index (DPI), whose pseudo-labels are generated by a large language model (LLM) teacher from clinical variables. A neural ordinary differential equation (ODE) models shared latent evolution along the disease-stage axis, while a separate nonnegative progression speed function links disease stage to chronological time under longitudinal supervision. Conditioned on the initial state, ODE prediction, patient context, and visit-level treatment information, a conditional diffusion model learns the residual distribution of individual latent changes beyond the shared progression trajectory. Together, these components establish a population-level progression prior with patient-specific stochastic refinement within a unified functional-imaging-based world model.

The main contributions are threefold: (1) a disease-stage parameterization that models shared metabolic progression independently of chronological time; (2) a nonnegative progression speed function that connects latent disease stage with chronological time; and (3) a conditional residual diffusion formulation that captures individual variation beyond the shared progression trajectory.

2 Related Works

Neural ODEs for disease progression. Neural ODEs provide a continuous formulation for hidden-state dynamics by parameterizing the derivative of a latent state with respect to an evolution variable Chen et al. (2018). ImageFlowNet extends neural ODE/SDE dynamics to multiscale medical-image forecasting, demonstrating their ability to model continuous changes in medical observations Liu et al. (2025). These approaches primarily formulate progression through continuous dynamics, while the progression variable is typically tied to temporal evolution. ADWM instead uses a weakly supervised disease progression index to define a disease-stage coordinate, with a separate nonnegative speed function mapping disease stage to chronological time when longitudinal supervision is available. This formulation distinguishes the direction of disease progression from the rate at which it unfolds across the individuals.

Diffusion-based progression modeling. Diffusion models provide flexible distributional modeling through iterative denoising Ho et al. (2020). In medical imaging, BrLP and Δ-Diffusion have applied diffusion formulations to longitudinal MRI and PET synthesis, respectively Puglisi et al. (2025); Sun et al. (2026). Closely related, cLDM-ODE integrates latent diffusion with neural ODEs for multimodal AD forecasting Sharma et al. (2026). Rather than using diffusion to directly model the complete longitudinal transition, ADWM uses it as a conditional residual model around the disease-stage ODE trajectory. The diffusion process is conditioned on the initial state, ODE prediction, patient context, and visit-level treatment information, allowing stochastic individual variation to be modeled separately from shared progression dynamics in a PET–SUVR latent space.

Medical world models. Recent medical world models have begun to model disease evolution and future observations under contextual conditions. MeWM simulates tumor evolution under treatment conditions Yang et al. (2025), while BrainWorld generates future fMRI conditioned on structural anatomy Xia et al. (2026). These works extend generative modeling from individual image synthesis toward condition-dependent simulation of evolving medical states. However, ADWM applies this world-model perspective to AD progression by representing PET and SUVR values in an aligned latent space and combining disease-stage dynamics with patient- and visit-specific conditions. In this setting, the ODE prior, progression-speed function, and residual diffusion jointly define a latent transition model, while SUVR provides quantitative readouts of the simulated metabolic state.

3 Methodology

ADWM models changes in brain metabolism along a disease progression index (DPI) through a shared progression model and a conditional distribution of individual deviations, as shown in Fig. 1. We encode paired PET imaging and regional SUVR values (SUVR reflects the metabolic intensity of the corresponding brain regions) into a joint latent state, establish a weakly supervised DPI

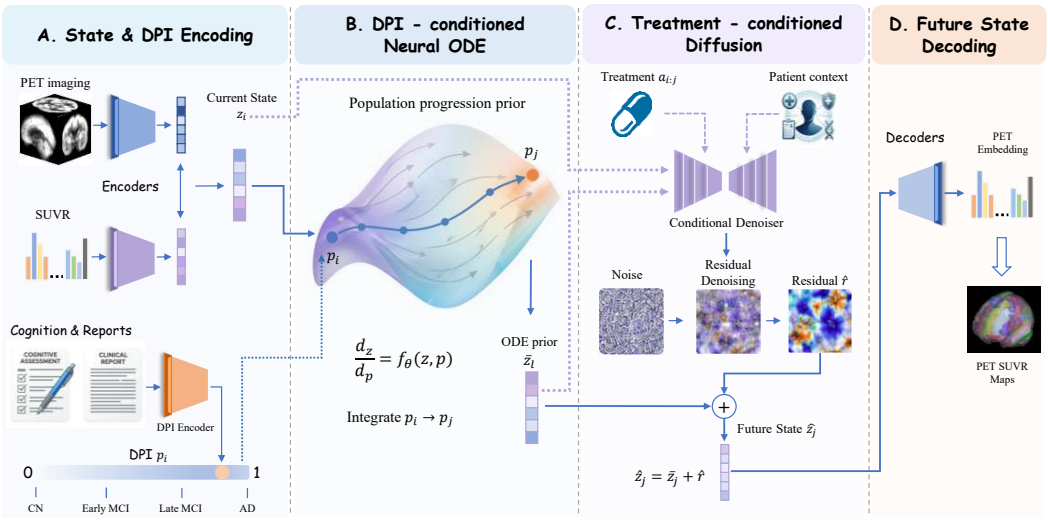

Figure 1: **Overview of ADWM.** ADWM encodes PET imaging and SUVR into a joint latent state, models disease progression with DPI-guided ODE dynamics, and refines the predicted state via conditional residual generation using patient context and treatment information.

coordinate, and learn a neural ODE along this coordinate. A conditional flow-matching model refines the ODE prediction using patient context and visit-level treatment information before regional metabolic readout. Implementation details are collected in Appendix B.

3.1 Problem Formulation

For patient n, let $\mathcal{D}_n = \{(x_{ni}, y_{ni}, \ell_{ni}, a_{ni})\}_{i=1}^{V_n}$ denote a visit sequence. Here, x_{ni} is a three-dimensional PET scan, $y_{ni} \in \mathbb{R}^R$ is a regional SUVR vector, ℓ_{ni} contains clinical evidence for weak stage supervision, and a_{ni} is a visit-level medication summary. We write c_{ni} for patient context available at the source visit and suppress the patient index for a single transition $i \to j$.

The model operates on a joint state $z_i \in \mathbb{R}^D$ and a scalar DPI coordinate $p_i \in [0, 1]$. Given a source state and a supplied target DPI $p_q \in [p_i, 1]$ within the modeled range, it constructs

$$\bar{z}_q = \mathcal{F}_p(z_i, p_i, p_q), \qquad r_q \sim P_\psi(\cdot \mid z_i, \bar{z}_q, c_i, a_i, a_q), \qquad \hat{z}_q = \bar{z}_q + r_q. \tag{1}$$

The target stage is an explicitly supplied input. Equation equation 1 defines a stochastic latent transition with a deterministic observation readout. Its residual distribution captures variation associated with patient attributes, recorded treatment, and shared-prior errors, without identifying these sources separately. Since treatment is observed rather than randomized, the model yields a conditional predictive distribution; changing a treatment input does not by itself identify a causal intervention effect.

3.2 Metabolic State Representation and Weak DPI Supervision

Joint PET–SUVR state. We adopt the pretrained ReMAP-PET representation (Dai et al., 2026), its PET and SUVR encoders produce aligned, individually normalized embeddings:

$$z_i^{\mathrm{P}} = \frac{E_{\mathrm{P}}(x_i)}{\|E_{\mathrm{P}}(x_i)\|_2}, \qquad z_i^{\mathrm{S}} = \frac{E_{\mathrm{S}}(y_i)}{\|E_{\mathrm{S}}(y_i)\|_2}, \qquad z_i = [z_i^{\mathrm{P}}; z_i^{\mathrm{S}}]. \tag{2}$$

The encoders remain fixed during dynamics and residual learning. The joint configuration uses $d_{\mathrm{P}} = d_{\mathrm{S}} = 256$, $D = 512$, and $R = 120$ regions. Retaining both embedding blocks preserves their aligned representations without assuming identical coordinates. Appendix B.1 specifies the SUVR-only variant and its corresponding loss blocks.

Clinical teacher and latent DPI readout. A clinical teacher assigns a pseudo-label $\tilde{p}_i \in [0, 1]$ from whitelisted evidence at visit i. The teacher returns a central score, an uncertainty interval,

supporting evidence, and an acceptance status. Only accepted labels receive a positive supervision weight w_i; other labels are masked from the DPI objective. DPI supplies an ordinally oriented stage coordinate rather than a calibrated probability of AD. We distill this supervision into a latent readout $p_i = h_\eta(z_i)$, implemented by an MLP with a sigmoid output. For weighted predictions and labels, define

$$\mathcal{H}_w(u, v) = \frac{\sum_k w_k \rho_\delta(u_k - v_k)}{\sum_k w_k + \epsilon}, \qquad \rho_\delta(e) = \begin{cases} \frac{1}{2}e^2, & |e| \le \delta, \\ \delta(|e| - \frac{1}{2}\delta), & |e| > \delta. \end{cases} \tag{3}$$

The anchor loss is $\mathcal{L}_{\text{anc}} = \mathcal{H}_w(h_\eta(z), \tilde{p})$, evaluated over both endpoints of a training batch. The readout is first warmed up on accepted training labels. During prior training, accepted teacher labels directly specify the integration coordinates; at inference, $p_i = h_\eta(z_i)$ initializes the source DPI. Clinical supervision details are given in Appendix B.1, and the readout architecture in Appendix B.3.1.

3.3 Shared DPI progression dynamics and prior learning

Disease-stage vector field. Following the continuous-state formulation of neural ODEs (Chen et al., 2018), we parameterize metabolic change per unit DPI by

$$\frac{dz}{dp} = f_\theta(z, p), \qquad z(p_i) = z_i. \tag{4}$$

All patients share f_θ, while their trajectories depend on their initial metabolic states. Sharing a vector field therefore does not force patients onto an identical curve. The stage-conditioned prior is

$$\bar{z}_j = \mathcal{F}_p(z_i, p_i, p_j) = z_i + \int_{p_i}^{p_j} f_\theta(z(p), p)\, dp. \tag{5}$$

Neither treatment variables nor a patient identity embedding enters this field. Its population interpretation follows from parameter sharing and pooled training. Since the training observations can include treated patients, the prior is not assumed to represent an untreated counterfactual trajectory.

Pair construction and training objective. Patients are partitioned before constructing visit pairs, so all visits from one patient belong to the same split. Prior training supports pooled stage-ordered pairs across patients, providing cross-sectional stage-transition supervision rather than observed within-patient transitions. For accepted labels with $\tilde{p}_j \ge \tilde{p}_i$, we initialize at $p_i = \tilde{p}_i$ and integrate to $p_j = \tilde{p}_j$ using differentiable fourth-order Runge–Kutta integration (Appendix B.3.1).

For a pair batch $\mathcal{B}$, the latent reconstruction objective is

$$\mathcal{L}_z = \frac{1}{2|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \left(\|\bar{z}_j^{\text{P}} - z_j^{\text{P}}\|_2^2 + \|\bar{z}_j^{\text{S}} - z_j^{\text{S}}\|_2^2 \right). \tag{6}$$

We also constrain the DPI readout at generated endpoints and penalize drift from the normalized encoder geometry:

$$\mathcal{L}_{\text{fut}} = \mathcal{H}_{w_j}(h_\eta(\bar{z}_j), \tilde{p}_j), \qquad \mathcal{L}_{\text{norm}} = \frac{1}{|\mathcal{B}|\,|\mathcal{M}|} \sum_{(i,j)\in\mathcal{B}} \sum_{m\in\mathcal{M}} \left(\|\bar{z}_j^m\|_2 - \|z_i^m\|_2 \right)^2, \tag{7}$$

where $\mathcal{M}$ is the set of retained modality blocks. The norm penalty discourages norm drift without enforcing exact unit norms. The integration interval fixes the target DPI, whereas $\mathcal{L}_{\text{fut}}$ supervises the readout at the generated metabolic state. The prior objective is

$$\begin{aligned} \mathcal{L}_{\text{stage}} = \mathcal{L}_z &+ \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{fut}}\mathcal{L}_{\text{fut}} + \lambda_{\text{norm}}\mathcal{L}_{\text{norm}} \\ &+ \lambda_{\text{cons}}\mathcal{L}_{\text{cons}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{ROI}}\mathcal{L}_{\text{ROI}}. \end{aligned} \tag{8}$$

The last three terms retain the auxiliary consistency, alignment, and regional-supervision penalties. The shared-prior architecture and integration details are given in Appendix B.3.1; reference settings are listed in Table 4.

3.4 CONDITIONAL GENERATION OF INDIVIDUAL RESIDUALS

After fitting and selecting the prior on validation patients, we freeze its parameters and construct residual targets from same-patient visit pairs:

$$r_{ij}^{\star} = z_j - \mathrm{sg}[\bar{z}_j], \qquad C_{ij} = [z_i; \bar{z}_j; c_i; a_i; a_j], \tag{9}$$

where sg denotes stop-gradient. Conditioning on the source state and prior endpoint allows the residual model to account for state-dependent prior errors. The context $c_i \in \mathbb{R}^{54}$ contains demographic, baseline clinical, and medical-history features; each treatment summary $a_i, a_j \in \mathbb{R}^{31}$ describes one visit. Their feature definitions and preprocessing are specified in Appendix B.2. The target state z_j enters only the residual training target and its noisy interpolation. Target DPI reaches the residual network only through $\bar{z}_j$, not as a separate scalar input.

Residual flow matching. We use conditional flow matching (Lipman et al., 2023) to transport a Gaussian reference distribution to the conditional residual distribution. Let $\tau \in [0, 1]$ denote the generative interpolation parameter, distinct from DPI p. For $\epsilon \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}(0, 1)$, define

$$r_0 = \sigma_r \epsilon, \qquad r_\tau = (1-\tau) r_0 + \tau r_{ij}^{\star}, \qquad u_\tau = r_{ij}^{\star} - r_0. \tag{10}$$

The reference scale σ_r is estimated from cached training residuals by Eq. equation 18. Noise and observed residuals are independently paired, without minibatch optimal-transport coupling. A network $v_\psi(r_\tau, \tau, C_{ij})$ predicts the path velocity using the weighted objective

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{(i,j),\tau,\epsilon}\left[\frac{\omega_{ij}}{D}\|v_\psi(r_\tau, \tau, C_{ij}) - u_\tau\|_2^2\right], \tag{11}$$

where $\omega_{ij} = \max(w_i, w_j)$. This weighting convention is distinct from the normalized weighted Huber loss used for DPI supervision.

Conditioning and residual regularization. We embed τ with sinusoidal features and project their concatenation with C_{ij} through an MLP. The resulting condition modulates a residual MLP through feature-wise affine transformations $h \mapsto (1+\gamma) \odot h + \beta$. Conditions therefore guide the residual stream without replacing it. The training endpoint estimator and complete residual objective are

$$\tilde{r}_1 = r_\tau + (1-\tau) v_\psi(r_\tau, \tau, C_{ij}), \tag{12}$$

$$\mathcal{L}_{\mathrm{res}} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{end}}\mathbb{E}\Big[\frac{\omega_{ij}}{D}\|\tilde{r}_1 - r_{ij}^{\star}\|_2^2\Big] + \lambda_{\mathrm{mag}}\mathbb{E}\Big[\frac{\omega_{ij}}{D}\|\tilde{r}_1\|_2^2\Big]. \tag{13}$$

By Eq. equation 19, the endpoint term reweights the velocity error by $(1-\tau)^2$ along the linear training path. The magnitude term shrinks corrections toward the shared prior and can reduce sample dispersion. Neither term guarantees calibrated uncertainty. The residual architecture and regularization analysis are provided in Appendix B.3.2, with hyperparameters in Table 4.

Sampling and treatment scenarios. At inference, we sample $r(0) \sim \mathcal{N}(0, \sigma_r^2 I)$ and solve

$$\frac{dr}{d\tau} = v_\psi(r, \tau, C_{iq}), \qquad \hat{z}_q = \bar{z}_q + r(1). \tag{14}$$

Residual sampling uses fixed-step Euler integration by default; solver options and step counts are given in Appendix B.3.2. A query treatment summary a_q must be supplied as a scenario. Using the recorded target-visit summary evaluates prediction conditional on that treatment.

3.5 REGIONAL METABOLIC READOUT AND INFERENCE

The generated joint state retains PET and SUVR embedding blocks. The ReMAP-PET SUVR head D_{S} was trained on normalized PET embeddings, giving the regional readout

$$\hat{y}_q^{(s)} = D_{\mathrm{S}}\Big(\mathrm{Norm}(\hat{z}_q^{\mathrm{P},(s)})\Big), \qquad s = 1, \ldots, S, \tag{15}$$

where Norm denotes ℓ_2 normalization. For $S > 1$, we compute the regional sample mean and componentwise variance using Eq. equation 20. Each sample is read out before aggregation because

the observation mapping is nonlinear. These statistics describe generator dispersion rather than automatically calibrated predictive intervals. An alternative jointly supervised readout G_ξ predicts standardized SUVR from the complete retained state, including the SUVR-only configuration; outputs are restored using training-set regional statistics (Appendix B.5). The quantitative observation output is regional SUVR.

Training caches fixed encoder states, warms up the DPI readout, fits the shared prior, and then learns residuals with that prior held fixed. Residual learning always uses same-patient visit pairs, including when prior training uses pooled pairs. Both stages are selected on validation patients, with test patients reserved for final evaluation. Freezing the prior fixes the residual targets without implying statistical independence between the components.

For a query, we encode the source observations, initialize $p_i = h_\eta(z_i)$, integrate to the supplied p_q, construct patient and treatment conditions, and generate independent residual samples before regional readout. The staged training and endpoint-simulation algorithms are specified in Appendix B.4. This is a stage-conditioned endpoint model; interpolating residuals across queried stages does not establish a learned stochastic path law.

4 EXPERIMENTS

First, we evaluate ADWM through ablations of flow matching (FM), treatment conditioning, patient context, and PET inputs. Our analysis examines whether improvements in state prediction translate into improvements in cognitive prediction, as reported in Table 1. In addition, we compare ADWM against four primary baseline methods as reported in Table 2. To ensure a fair comparison across different modeling paradigms, all models generate longitudinal trajectories conditioned on baseline covariates and are evaluated at the subjects' actual observed timestamps.

4.1 EXPERIMENTAL SETUP

Materials. We use longitudinal fluorodeoxyglucose (FDG)-PET data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Jack et al. (2015) that measures cerebral glucose metabolism for assessing dementia, comprising 1,489 PET scans from 658 subjects, including cognitively normal (CN), mild cognitive impairment (MCI), and AD stage. Each subject has PET scans acquired at multiple time points. Detailed demographic and clinical info and preprocessing are provided in C.1.

Task and settings. Given an initial disease state and the conditioning information available under the task definition, ADWM predicts the state at a target disease stage or follow-up visit. Detailed experimental settings, including dataset splits, prediction horizons, conditioning variables, training configurations, and inference settings, are provided in C.2.

Evaluation metrics. In this study, we evaluate disease progression from four aspects: latent trajectory fidelity, biomarker fidelity, clinical fidelity, and distributional fidelity. All metrics are evaluated at the subjects' actual observation timestamps and aggregated across prediction horizons. Detailed definitions and interpretations of these metrics are provided in C.3.

Baseline methods. We compare ADWM with four baselines reported in Table 2: LSTM Wang et al. (2018), Latent ODE Rubanova et al. (2019), Conditional Diffusion Xiao et al. (2024), and AlzheimerODE Jiang et al. (2026). These baselines represent discrete-time sequence modeling, continuous-time latent dynamics, conditional diffusion generation, and ODE-based AD progression modeling, respectively. Implementation details and hyperparameter settings of ADWM are provided in Appendix A.2.

4.2 ABLATION STUDY

State prediction improves over persistence. In Table 1, the full ADWM achieves PET and SUVR CosSim of 0.969 and 0.789, respectively, compared with 0.881 and 0.296 for Persistence. SUVR MAE decreases from 0.151855 to 0.093609, a relative reduction of 38.4%, while ΔSUVR MAE decreases from 0.151 to 0.087, a reduction of 42.0%. NRMSE and RMSE decrease by 47.3% and

Table 1: Ablation results of ADWM for disease-state and cognitive-score prediction. Δ values indicate the difference from the full ADWM.

Configuration	PET CosSim ↑	SUVR CosSim ↑	SUVR MAE ↓	ΔSUVR MAE ↓	MMSE MAE ↓	NRMSE ↓	RMSE ↓
ADWM	**0.969**	**0.789**	**0.094**	**0.087**	**1.991**	**0.053**	**0.030**
Ablation variants							
w/o FM	0.950	0.670	0.102	0.103	1.687	0.062	0.035
Δ	−0.019	−0.119	+0.008	+0.016	−0.304	+0.009	+0.005
w/o Treatment	0.969	0.796	0.094	0.090	2.103	0.053	0.030
Δ	0.000	+0.007	0.000	+0.003	+0.112	0.000	0.000
w/o Context	0.969	0.795	0.091	0.089	2.121	0.052	0.030
Δ	0.000	+0.006	−0.003	+0.002	+0.130	−0.001	0.000
Only SUVR (FM)	–	0.789	–	–	2.495	0.076	0.040
Δ	–	0.000	–	–	+0.504	+0.023	+0.010
Only SUVR (ODE)	–	0.669	–	–	1.985	0.107	0.055
Δ	–	−0.120	–	–	−0.006	+0.054	+0.025
Persistence	0.881	0.296	0.152	0.151	1.902	0.101	0.057
Δ	−0.088	−0.493	+0.058	+0.064	−0.089	+0.048	+0.027

46.7%, respectively. These results support the value of predicting state evolution over carrying the initial state forward on the reported state-related metrics. This benefit does not extend to MMSE: the full model obtains an MAE of 1.991, compared with 1.902 for Persistence. Improved metabolic-state prediction therefore does not automatically imply improved cognitive prediction.

FM improves state-related metrics but increases MMSE error. Adding FM to DPI-ODE increases PET CosSim from 0.950 to 0.969 and SUVR CosSim from 0.670 to 0.789, corresponding to absolute gains of 0.019 and 0.119. SUVR MAE and ΔSUVR MAE decrease by 8.7% and 15.0%, respectively, while NRMSE and RMSE both decrease by approximately 14.1%. These observations are consistent with the intended role of FM as a correction to the ODE prediction. The larger relative reduction in ΔSUVR MAE concerns that metric's specific evaluation target; it does not by itself establish accurate long-term trajectories or progression rates.

The same modification increases MMSE MAE from 1.687 to 1.991, or approximately 18.0%. DPI-ODE without FM achieves the lowest MMSE MAE among the reported configurations. Thus, the evidence supports a benefit for state prediction rather than a uniform benefit across downstream outcomes. One possible explanation is that the state correction and the MMSE predictor depend on different aspects of the representation, but testing this explanation requires further analysis of the readout and training objectives. Moreover, the absence of an FM-only configuration prevents this ablation from isolating the necessity of the DPI-ODE component in the full model.

Treatment conditioning provides selective benefits. Relative to the variant without treatment conditioning, the full model reduces SUVR MAE from 0.094 to 0.094, ΔSUVR MAE from 0.090 to 0.087, and MMSE MAE from 2.104 to 1.991. These correspond to reductions of 0.6%, 3.0%, and 5.3%, respectively. PET CosSim differs by only 0.000002. Conversely, removing treatment conditioning yields a higher SUVR CosSim (0.796 versus 0.789) and slightly lower NRMSE and RMSE. Treatment information therefore benefits selected quantitative targets rather than every reported metric. This ablation assesses the predictive utility of observed treatment information; it does not identify causal treatment effects.

Patient context introduces a target-dependent trade-off. Compared with the variant without patient context, the full model reduces ΔSUVR MAE from 0.0889 to 0.087 and MMSE MAE from 2.121 to 1.991, corresponding to reductions of 1.6% and 6.1%. However, the context-free variant achieves lower SUVR MAE (0.091 versus 0.094), NRMSE (0.052 versus 0.053), and RMSE (0.0296 versus 0.0302), as well as slightly higher PET CosSim. The benefits of context are therefore concentrated in the reported change target and cognitive outcome, without consistent improvements in state error. In the absence of repeated runs or confidence intervals, these small differences remain descriptive observations.

The contribution of PET depends on the evaluation target. With DPI-ODE and FM retained, the full and SUVR-only configurations have similar SUVR CosSim (0.789 and 0.789). Neverthe-

Table 2: Prediction results of ADWM and compared methods under different evaluation protocols.

Model	NRMSE ↓	RMSE ↓	JSD ↓	WD ↓	ADE ↓	CosSim ↑
LSTM	0.095	0.079	0.0274	2.180	1.860	0.868
Latent ODE	0.069	0.055	0.0126	1.240	1.060	0.890
Conditional Diffusion	0.082	0.068	0.0198	1.710	1.480	0.879
AlzheimerODE	0.074	0.246	0.0060	4.181	4.781	0.899
ADWM (Ours)	0.053	0.030	0.0001	0.578	0.427	0.876

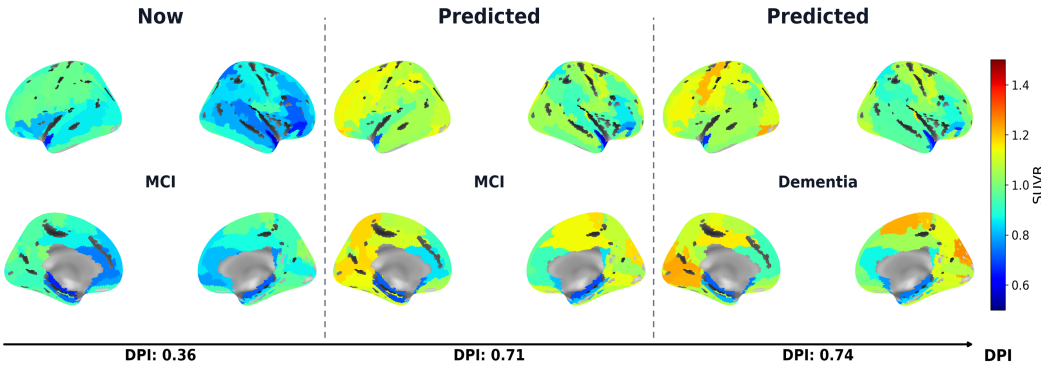

Figure 2: Predicted disease progression: Predicted longitudinal changes in the brain PET SUVR map along disease progression.

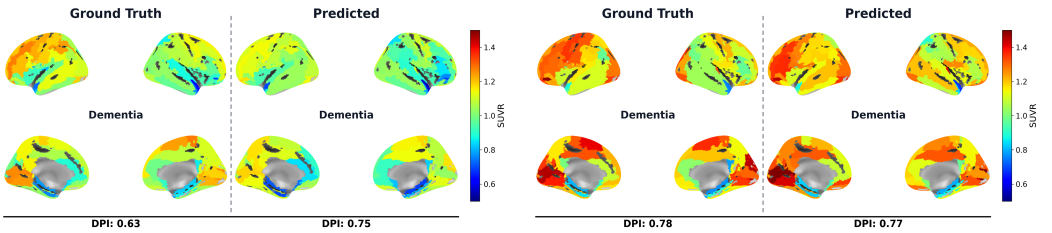

Figure 3: Predicted vs. ground truth: Comparison of the predicted and ground-truth brain PET SUVR maps at the future time point.

less, the full model reduces MMSE MAE from 2.495 to 1.991, a reduction of 20.2%. The reported NRMSE and RMSE are also 30.1% and 24.5% lower; these differences can be interpreted as performance gains only if both configurations are evaluated on the same targets and scales. Without FM, the joint-input DPI-ODE also has a lower MMSE MAE than its SUVR-only counterpart (1.687 versus 1.985). Together, these observations suggest that SUVR cosine similarity alone does not capture the predictive contribution of PET inputs.

Within the SUVR-only setting, adding FM increases SUVR CosSim from 0.669 to 0.789 and reduces NRMSE and RMSE by approximately 29.0% and 27.7%. MMSE MAE increases from 1.985 to 2.495. The state–cognition trade-off associated with FM is therefore observed in both input configurations. Additional readout and objective ablations are needed to determine its source.

4.3 AlzheimerODE Reference Results under Different Protocols

Table 2 reports the results of ADWM and AlzheimerODE under their respective evaluation settings. ADWM obtains NRMSE, RMSE, JSD, WD, ADE, and CosSim values of 0.053, 0.030, 0.0001, 0.578, 0.427, and 0.876. The corresponding reported values for AlzheimerODE are 0.074, 0.246, 0.0060, 4.181, 4.781, and 0.899. The prediction spaces, test samples, and normalization rules differ between the two protocols. Consequently, these values describe separate evaluations and do not constitute a controlled comparison.

Protocol differences affect every reported metric. RMSE, WD, and ADE depend on the scale and geometry of the prediction representation, while NRMSE also depends on its normalization ref-

erence. JSD depends on the construction and support of the evaluated distributions. Although standard cosine similarity is invariant to positive rescaling of either vector, it remains sensitive to the representation, preprocessing, and sample composition. Neither the smaller numerical errors and distances reported for ADWM nor the higher CosSim reported for AlzheimerODE therefore establish a methodological advantage.

A direct comparison requires shared observable prediction targets, test subjects, forecast horizons, information-availability constraints, normalization, and metric implementations. Models with different internal latent spaces should be evaluated after mapping their predictions to shared observable targets. The present evidence for the contribution of ADWM's components comes primarily from the internal ablations; comparative superiority over AlzheimerODE remains untested under a unified protocol.

4.4 Visual Interpretability

In the visual analysis, we visualize the predicted future metabolic states to examine the validity of ADWM and its ability to model disease progression. As shown in Fig. 2, the predicted metabolic changes for an individual progressing toward MCI or AD are consistent with the expected pattern of AD progression. In particular, pronounced metabolic changes are observed in regions including the parietal and frontal lobes, indicating that the model captures spatially meaningful metabolic alterations associated with disease progression. We further compare the predicted and ground-truth SUVR maps as shown in Fig. 3, the predicted maps exhibit highly similar spatial distribution patterns to the ground-truth maps, indicating that ADWM captures the spatial characteristics of metabolic alterations. These results demonstrate that ADWM can capture both the spatial characteristics and longitudinal changes of metabolic patterns during disease progression. Detailed slice-level visualizations of the corresponding metabolic changes are provided in C.4.

4.5 Discussion

The ablations reveal distinct contributions from ADWM's components. FM consistently improves several state-related metrics, whereas the benefits of treatment and patient context concentrate on the change target and MMSE. The full model achieves the lowest reported ΔSUVR MAE, while DPI-ODE without FM achieves the lowest MMSE MAE. These findings motivate evaluating ADWM against specific prediction objectives rather than summarizing its behavior with a single similarity score. The supplied results do not include run-to-run variability, patient-level confidence intervals, horizon-stratified evaluations, or external-cohort validation. Our conclusions are therefore restricted to the reported point estimates. Establishing cross-method superiority additionally requires evaluation under a common protocol.

5 Conclusion

We presented ADWM, a latent world model for AD progression from sparse longitudinal functional imaging. ADWM combines aligned PET–SUVR representations with a weakly supervised disease progression index to model shared metabolic dynamics along a continuous disease-stage coordinate. A separate progression-speed function connects disease stage to chronological time under longitudinal supervision, while conditional residual flow matching captures individual variation around the shared ODE prediction using patient context and treatment information. This decomposition links population-level dynamics with individualized stochastic predictions, with regional SUVR readouts providing a quantitative interpretation of simulated states. The framework supports studying heterogeneous metabolic progression within a common latent state space. Future work will focus on prospective validation, uncertainty calibration, and identifying treatment effects beyond observational associations.

REFERENCES

Maryamossadat Aghili, Solale Tabarestani, Malek Adjouadi, and Ehsan Adeli. Predictive modeling of longitudinal data for Alzheimer's disease diagnosis using RNNs. In *Predictive Intelligence in Medicine (PRIME 2018)*, Lecture Notes in Computer Science, pp. 112–119. Springer, Springer International Publishing, 2018. doi: 10.1007/978-3-030-00320-3_14.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583, 2018.

Dasen Dai, Yanteng Zhang, Shuoqi Li, Yuxiang Wei, Hongjie Yu, Qingxin Zhang, Qizhen Lan, Jagath C. Rajapakse, and Vince D. Calhoun. Remap-pet: Beyond visual understanding–learning region-guided metabolic alignment semantics from brain pet. *arXiv preprint arXiv:2606.29577*, 2026. doi: 10.48550/arXiv.2606.29577.

Bhoomishree Gowda, Adithi Ambatipudi, Aashna Ali, and Shruthi M L J. Hippocampal time-flow modeling: A neural ode framework for Alzheimer's forecasting. In *2026 IEEE 4th International Conference on Device Intelligence, Computing and Communication Technologies (DICCT)*, pp. 27–32. IEEE, IEEE, 2026. doi: 10.1109/DICCT69099.2026.11535969.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning*, pp. 2555–2565. PMLR, PMLR, 2019.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Clifford R. Jack, Josephine Barnes, Matt A. Bernstein, Bret J. Borowski, James Brewer, Shona Clegg, Anders M. Dale, Owen Carmichael, Christopher Ching, Charles DeCarli, et al. Magnetic resonance imaging in Alzheimer's disease neuroimaging initiative 2. *Alzheimer's & Dementia*, 11(7):740–756, 2015. doi: 10.1016/j.jalz.2015.05.002.

Yuhao Jiang, Junbo Ma, Guoqiu Wen, Xinjie Han, Hongqing He, Jiayang Su, Fan Yang, and Minghan Chen. Alzheimerode: A knowledge-guided generative model for Alzheimer's disease progression. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 11201–11212. ACM, 2026. doi: 10.1145/3770855.3818833.

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

Chen Liu, Ke Xu, Liangbo L. Shen, Guillaume Huguet, Zilong Wang, Alexander Tong, Danilo Bzdok, Jay Stewart, Jay C. Wang, Lucian V. Del Priore, et al. ImageFlowNet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10890535.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. Brain latent progression: Individual-based spatiotemporal disease progression on 3D brain MRIs via latent diffusion. *Medical Image Analysis*, 106:103734, 2025. ISSN 1361-8415. doi: 10.1016/j.media.2025.103734.

Yulia Rubanova, Ricky T. Q. Chen, and David K. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, pp. 342–347, 2019.

Rishabh Sharma, Vinay Kukreja, Shanmugasundaram Hariharan, and Shih-Yu Chen. cLDM-ODE: A multimodal generative framework for uncertainty-aware forecasting of Alzheimer's disease progression. *IRBM*, 47(1):100926, 2026. ISSN 1959-0318. doi: 10.1016/j.irbm.2025.100926.

Yongheng Sun, Minhui Yu, Mengqi Wu, Maureen Kohi, and Mingxia Liu. Delta-diffusion: Modeling longitudinal brain amyloid-PET trajectories via conditional poisson diffusion bridge. *arXiv preprint arXiv:2606.22216*, 2026. doi: 10.48550/arXiv.2606.22216.

Tingyan Wang, Robin G. Qiu, and Ming Yu. Predictive modeling of the progression of Alzheimer's disease with recurrent neural networks. *Scientific Reports*, 8(1):9161, 2018. doi: 10.1038/s41598-018-27337-w.

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*, 2026. doi: 10.48550/arXiv.2606.17742.

Qing Xiao, Siyeop Yoon, Hui Ren, Matthew Tivnan, Lichao Sun, Quanzheng Li, Tianming Liu, Yu Zhang, and Xiang Li. Conditional score-based diffusion model for cortical thickness trajectory prediction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 78–87, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72069-7. doi: 10.1007/978-3-031-72069-7_8.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model: Generative simulation of tumor evolution for treatment planning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. doi: 10.1109/ICCV51701.2025.00779.

A Baselines and Implementation Details

A.1 Baseline Descriptions

All baselines are grouped into four families that correspond to the categories listed in Section 4, and each family contributes one representative to the main comparison. Every baseline is trained and evaluated on the same cached ADWM latent states and the same patient-level split used by ADWM itself, so that differences reflect the transition model rather than the representation.

Discrete-time sequence models (LSTM/GRU). Recurrent progression models encode a patient's follow-up history and learn a discrete-time state transition, forecasting the next state from the encoded sequence (Aghili et al., 2018; Wang et al., 2018). They represent the traditional deep sequence-modeling approach for longitudinal AD data and test whether a continuous dynamical formulation is preferable to a discrete recurrence when visits are irregularly spaced.

Continuous-time latent dynamics (Latent ODE). Latent ODE models place the transition in a latent space and parameterize $\mathrm{d}z/\mathrm{d}t = f(z, t)$, which suits sparse and irregularly sampled series and can in principle capture a shared population trajectory (Rubanova et al., 2019; Gowda et al., 2026). This family is the central control for the shared prior: comparing it against ADWM isolates the contribution of the disease-stage coordinate and of the residual generator, since it contains a continuous prior but neither of those two components.

Conditional generative models (conditional diffusion). Conditional score-based diffusion predicts future medical states by learning $p(x_{\text{future}} \mid \text{condition})$ and sampling through iterative denoising (Xiao et al., 2024). This family represents the generative alternative and tests whether the disease-stage ODE prior adds value beyond a purely conditional generator, which is also the closest comparison for the residual module of ADWM.

Latent state-space world models (RSSM/Dreamer). Recurrent state-space models learn a transition $p(s_{t+1} \mid s_t, a_t)$ over latent states and roll the state forward under actions (Hafner et al., 2020; 2019). Mapping the latent state to the joint PET–SUVR embedding and the visit-level treatment summary to the action gives the generic world-model baseline, which tests whether the medical state-space formulation of ADWM is preferable to a general-purpose agent world model.

Related advanced method (cLDM-ODE). cLDM-ODE combines latent diffusion with a neural ODE for multimodal AD progression forecasting (**?**). It is the closest prior work, but it is included as a related-work comparison rather than a primary baseline: it does not expose a treatment-action interface, and it does not separate a shared progression trend from individual variation in the way ADWM does.

A.2 Implementation Details

The full architecture, numerical-integration scheme, and reference hyperparameters are reported in Appendix B.3; the staged training procedure and the endpoint-simulation routine are given in Appendix B. The prompt specification used to obtain the clinical pseudo-labels is given in Appendix D. This subsection records only the shared implementation contract that the baselines inherit from ADWM.

Shared inputs. Every method consumes the same frozen encoders and the same cached states: per-visit PET and SUVR embeddings from the pretrained ReMAP-PET representation, the accepted visit-level DPI pseudo-labels and their weights, the 54-dimensional patient context, and the 31-dimensional visit-level treatment summaries. No baseline retrains or fine-tunes an encoder.

Shared protocol. Patients are partitioned before any visit pair is constructed, so all visits of a patient lie in a single split. Models are selected on validation patients and evaluated once on held-out test patients at their actually observed visit timestamps, with metrics aggregated across prediction horizons. Where a baseline produces samples rather than a point estimate, the reported value is the sample mean unless the metric is explicitly a distributional one.

Symbol	Definition
$p, \tilde{p}$	DPI coordinate and accepted clinical-teacher pseudo-label.
p_q	Supplied target DPI for an endpoint query.
τ	Generative interpolation parameter in $[0, 1]$, distinct from DPI.
$z = [z^{\mathrm{P}}; z^{\mathrm{S}}]$	Joint PET–SUVR state with two 256-dimensional blocks.
h_η, f_θ	Latent-to-DPI readout and metabolic change per unit DPI.
$\bar{z}, \hat{z}$	Shared-prior endpoint and residual-refined state.
c_i, a_i, a_j	Source-visit patient context and two visit-level treatment summaries.
$r^\star, r_\tau, v_\psi$	Residual target, interpolated residual, and conditional velocity network.
D_{S}, G_ξ	Pretrained PET-to-SUVR head and optional joint-latent SUVR readout.

Table 3: Notation for DPI-conditioned metabolic endpoint generation.

Reported quantities. Ablation results in Table 1 are single point estimates; run-to-run variability is not reported and no confidence intervals are claimed. Reference results for ADWM and AlzheimerODE in Table 2 were produced under separate evaluation protocols, and the protocol differences are discussed in Section 4.3.

B Additional Methodological Details

This appendix supplements the DPI-axis formulation in Section 3 with representation variants, conditioning specifications, network and solver settings, training and simulation algorithms, and predictive summaries. The progression equations and learning objectives are given in the main text.

B.1 Representation variants and clinical DPI supervision

The joint configuration uses the fixed PET and SUVR encoders in Eq. equation 2, with two individually normalized 256-dimensional embedding blocks. A SUVR-only configuration retains $z_i = z_i^{\mathrm{S}}$ and $D = 256$. In that configuration, only the SUVR block contributes to Eq. equation 6, and the modality set in Eq. equation 7 is $\mathcal{M} = \{\mathrm{S}\}$. The encoders remain fixed throughout prior and residual learning.

The clinical teacher uses whitelisted evidence from each visit and returns a central stage score, an uncertainty interval, supporting evidence, and an acceptance status. Only accepted labels receive positive supervision weights. The weighted Huber anchor loss is evaluated over both endpoints of each training batch. Accepted teacher labels determine the source and target coordinates for prior training, whereas the learned metabolic readout h_η estimates the source DPI for a new query. The complete system prompt, user template, and provider-assembly contract are specified in Appendix D, together with the distinction between machine-checked constraints and prompt-level guidance.

B.2 Notation and conditioning contract

The patient context has 54 entries: 15 numerical attributes, 20 one-hot categorical entries, and 19 medical-history indicators. Numerical values use median/IQR normalization, with unavailable numerical values represented by zero after normalization. The medication encoder has 31 entries: six AD-drug dose summaries, fifteen other medication indicators, one vitamin indicator, six PET-impact categories, a normalized medication count, a total AD-drug dose summary, and an AD-medication indicator. These are engineered features rather than learned patient-identity embeddings. Patient context must be available at the source visit, and a_q is supplied as the query treatment scenario.

The residual network receives $C_{ij} = [z_i; \bar{z}_j; c_i; a_i; a_j]$ and an embedding of τ. The target DPI is not supplied as a separate scalar; its influence is mediated by $\bar{z}_j$. Consequently, different target DPIs that yield the same prior endpoint and the same remaining conditions cannot be distinguished by this interface. The target metabolic state z_j is used only to construct the residual target and its training interpolation.

B.3 Network architecture, numerical integration, and configuration

B.3.1 Shared prior and DPI integration

For the joint configuration, f_θ is a $513 \to 512 \to 512 \to 512$ MLP whose input concatenates z and p, with tanh activations after the first two layers. The DPI readout is $512 \to 128 \to 64 \to 1$, with tanh hidden activations and a sigmoid output.

The prior is trained over accepted teacher-label intervals using Eq. equation 8. Because the integration interval fixes the target DPI, a separate penalty on the integrated stage endpoint is redundant. The term $\mathcal{L}_{\mathrm{fut}}$ remains informative because it supervises the readout $h_\eta(\bar{z}_j)$ at the generated metabolic endpoint. The norm penalty discourages drift from the normalized encoder geometry without enforcing exact unit norms.

For a current state (z_k, p_k) and DPI step h, differentiable fourth-order Runge–Kutta integration computes

$$\begin{aligned} k_1 &= f_\theta(z_k, p_k), \\ k_2 &= f_\theta(z_k + \tfrac{h}{2}k_1, p_k + \tfrac{h}{2}), \\ k_3 &= f_\theta(z_k + \tfrac{h}{2}k_2, p_k + \tfrac{h}{2}), \\ k_4 &= f_\theta(z_k + hk_3, p_k + h), \\ z_{k+1} &= z_k + \tfrac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4), \qquad p_{k+1} = p_k + h. \end{aligned} \tag{16}$$

For a batch of nonnegative DPI intervals Δp_b, the implementation uses $K = \max(1, \lceil \max_b \Delta p_b / h_{\max} \rceil)$ steps with per-sample step size $h_b = \Delta p_b / K$. Gradients pass through these explicit operations; an adjoint solver is not used. Integration applies within the modeled DPI range. Clipping returned DPI values to $[0, 1]$ is a numerical operation rather than a learned boundary condition.

An optional tangent projection replaces a modality-block velocity v^m by

$$v^m_{\tan} = v^m - \frac{\langle z^m, v^m \rangle}{\max(\|z^m\|_2^2, \epsilon)} z^m. \tag{17}$$

When $\|z^m\|_2^2 \geq \epsilon$, this velocity is orthogonal to z^m and preserves its norm in the exact continuous dynamics, subject to numerical integration error. The projection is optional and is not enabled in the supplied main launch configuration.

B.3.2 Residual generator and regularization

The residual network embeds τ with 128-dimensional sinusoidal features. Concatenating this embedding with C_{ij} gives $128 + 2(512) + 54 + 2(31) = 1268$ conditioning dimensions. A $1268 \to 512 \to 512$ MLP supplies the feature-wise affine modulation condition. Each of six residual blocks contains layer normalization, two linear maps, a SiLU activation, affine modulation, and dropout. The final layer is zero initialized. This initializes the velocity field to zero; the corresponding untrained flow leaves the sampled Gaussian residual unchanged rather than returning an identically zero residual.

The reference scale is estimated from a subset of cached training residuals as

$$\sigma_r = \max\big(\mathrm{Std}(r^\star), \mathrm{RMS}(r^\star), 10^{-3}\big). \tag{18}$$

Gaussian noise and observed residuals are independently paired for Eq. equation 10, without minibatch optimal-transport coupling. Combining the linear interpolation with Eq. equation 12 yields

$$\tilde{r}_1 - r^\star_{ij} = (1 - \tau)\big(v_\psi(r_\tau, \tau, C_{ij}) - u_\tau\big). \tag{19}$$

Thus, the endpoint penalty reweights the velocity error by $(1 - \tau)^2$. The magnitude penalty shrinks corrections toward the shared prior and can reduce sample dispersion. Neither penalty guarantees calibrated uncertainty. Distribution discrepancies computed from detached samples are monitoring quantities rather than gradient-based objectives.

Residual sampling uses a fixed-step solver over $\tau \in [0, 1]$. Euler integration is the default. The optional Heun solver applies a predictor–corrector update on nonfinal steps and an Euler update at the final step. A noise-prediction diffusion formulation is an alternative with the same state and conditioning interface; the main model uses flow matching.

Quantity	Supplied configuration
Joint latent / context / per-visit treatment dimensions	$512/54/31$
Residual hidden width / blocks / dropout	$512/6/0.1$
Generative interpolation embedding dimension	128
Residual optimizer and learning rate	AdamW, 3×10^{-4}
Residual weight decay and gradient clipping	10^{-5}, norm 1.0
Residual endpoint / magnitude weights	$1.0/0.05$
Residual training epochs	200
Residual batch size, joint DPI configuration	128
DPI-axis ODE maximum step	0.05
Residual validation / default inference steps	$10/20$
Eligible-patient split and random seed	$70/15/15\%$, seed 42

Table 4: Reference settings for the joint-state DPI configuration.

B.4 Training and DPI-conditioned endpoint simulation

Algorithm 1: staged learning.

1. Partition eligible patients before pair construction and cache the fixed PET–SUVR states. Attach accepted visit-level DPI labels and their supervision weights.
2. Warm up h_η using $\mathcal{L}_{\text{anc}}$.
3. Construct stage-ordered prior-training pairs, allowing pooled pairs across patients. Integrate between accepted source and target DPIs and optimize $\mathcal{L}_{\text{stage}}$ in Eq. equation 8.
4. Select the prior using validation patients. Freeze it and cache prior endpoints for same-patient residual-training pairs. These residual pairs are required even when the prior was trained on pooled pairs.
5. Draw a pair, Gaussian reference noise, and τ. Form $r^\star_{ij}$, C_{ij}, and r_τ using Eqs. equation 9 and equation 10, then update the residual model using $\mathcal{L}_{\text{res}}$ in Eq. equation 13.
6. Select the residual model using validation patients and reserve test patients for final evaluation.

Freezing the prior fixes the residual targets during generative training. This separates the optimization stages without implying statistical independence between the shared and residual components.

Algorithm 2: DPI-conditioned endpoint simulation.

1. Encode source PET and SUVR into z_i and initialize $p_i = h_\eta(z_i)$.
2. Supply a target DPI $p_q \geq p_i$ within the modeled range and integrate the shared field to obtain $\bar{z}_q = \mathcal{F}_p(z_i, p_i, p_q)$.
3. Construct C_{iq} using source-visit context and the specified treatment scenario. Draw S independent Gaussian residual initializations and integrate Eq. equation 14 for each draw.
4. Add each residual to the prior endpoint, apply a compatible regional readout, and return the resulting samples and sample summaries.

Using recorded target-visit treatment evaluates prediction conditional on that recorded scenario. The multi-target convenience wrapper that scales one sampled endpoint residual linearly over target indices is a visualization heuristic. It does not define a learned joint stochastic path; Algorithm 2 describes the trained endpoint model.

B.5 Regional readout and predictive summaries

The pretrained SUVR head uses the normalized PET embedding block as in Eq. equation 15. When raw regional measurements are available, an alternative jointly supervised readout $G_\xi : \mathbb{R}^D \to \mathbb{R}^R$ predicts standardized SUVR from the complete latent state. For the joint configuration, this is a

$512 \to 256 \to 120$ MLP with a tanh hidden layer. Its outputs are transformed back using training-set regional statistics. The alternative readout also supports the SUVR-only configuration, whose missing PET block precludes direct use of the PET-trained head.

For $S > 1$ regional predictions, the sample mean and componentwise variance are

$$\mu_y = \frac{1}{S}\sum_{s=1}^{S} \hat{y}_q^{(s)}, \qquad v_y = \frac{1}{S-1}\sum_{s=1}^{S} (\hat{y}_q^{(s)} - \mu_y)^{\odot 2}. \tag{20}$$

The same summaries can be computed in latent space. Because the regional readout is nonlinear, every sample is decoded before aggregation. These statistics describe dispersion under the learned generator and do not automatically provide calibrated predictive intervals. The implemented PET interface retrieves reference images using latent similarity; it does not synthesize new three-dimensional PET volumes.

C Experimental setup and additional results

All PET images are preprocessed using the standard pipeline, including spatial alignment to the corresponding MRI and normalization to MNI152 space. After preprocessing, PET images have a spatial resolution of $112 \times 128 \times 112$ voxels.

C.1 Dataset demographics, clinical characteristics, and preprocessing

Cohort. We use longitudinal FDG-PET from the Alzheimer's Disease Neuroimaging Initiative (ADNI), curated into the `dataset_ODE` release used by ADWM. After aligning the clinical table, FDG-PET volumes, and regional SUVR profiles by a canonical visit key , the cohort contains **2,584** matched visits from **1,123** subjects spanning cognitively normal (CN), mild cognitive impairment (MCI; EMCI+LMCI), and Alzheimer's disease (AD) stages. Each visit provides one FDG-PET scan (`.nii.gz`), one 120-region SUVR vector (extracted by AAL2 brain atlas), and the corresponding clinical row in a main table. Visits are scheduled at ADNI time points `bl`, `m06`, ..., `m96` (up to 96 months from baseline). Of the 1,123 subjects, 658 have $\geq$2 visits (longitudinal pool used for ODE/pair training); 465 singleton subjects are retained in the curated table for encoder/label coverage but are excluded from the train/val/test split (§C.2). And

Demographics and clinical scores. Table 5 summarises baseline demographics and clinical measures by baseline diagnosis (subject-level; first available visit in the curated table). MCI merges EMCI (N=275) and LMCI (N=334). APOE4+ denotes $\geq$1 ε4 allele. Continuous values are mean $\pm$ SD.

Table 5: Baseline demographics and clinical characteristics of the ADWM datasets cohort (subject-level). CN / MCI / AD are merged baseline stages; Overall is all subjects.

Characteristic	Overall	CN	MCI	AD
Subjects, N (%)	1123 (100)	296 (26.4)	609 (54.2)	218 (19.4)
Male, N (%)	631 (56.2)	147 (49.7)	356 (58.5)	128 (58.7)
Age, years	73.4 ± 7.3	74.3 ± 5.8	72.5 ± 7.6	74.8 ± 7.7
Education, years	16.0 ± 2.8	16.4 ± 2.8	16.0 ± 2.7	15.4 ± 2.9
APOE4+, N (%)	496 (44.2)	74 (25.0)	278 (45.7)	144 (66.1)
MMSE	27.1 ± 2.9	29.0 ± 1.3	27.6 ± 2.4	23.2 ± 2.2
CDRSB	1.78 ± 1.94	0.05 ± 0.25	1.62 ± 1.28	4.56 ± 1.66
FAQ	4.3 ± 6.5	0.2 ± 0.9	3.1 ± 4.4	13.5 ± 7.1
ADAS13	16.8 ± 9.9	9.4 ± 4.3	15.7 ± 7.5	30.4 ± 8.2
FDG (table summary)	1.23 ± 0.16	1.30 ± 0.11	1.25 ± 0.14	1.07 ± 0.14
DPI score	0.35 ± 0.23	0.13 ± 0.09	0.33 ± 0.14	0.70 ± 0.13
Visits / subject (mean)	2.30	2.45	2.38	1.88

Disease Progression Index (DPI). Each visit is annotated with a continuous DPI in $[0, 1]$ by an LLM teacher that reads the visit-level clinical payload (cognition, function, CSF/PET biomarkers, demographics; columns listed in the Stage-2 filter config) and returns a score with confidence and free-text rationale. Accepted labels enter training with a weak-supervision weight (`dpi_weight`$=0.5$ by default); strength is further controlled by Stage-2 loss weights λ_{dpi} / $\lambda_{\text{future_dpi}}$. DPI is the primary disease-stage axis for EXP2 experiments (§C.2).

Imaging preprocessing and latent encoding. FDG-PET volumes are encoded with a frozen ReMAP-PET foundation model (MedicalNet 3D ResNet-50 PET encoder with region-guided SUVR supervision): (i) percentile intensity clipping at $[0.5, 99.5]$ and rescaling to $[0, 1]$; (ii) trilinear resampling to 96^3; (iii) encoder + projector yielding a 256-d L2-normalised PET embedding z^{pet}. Regional SUVR is computed on the AAL2 atlas (120 grey-matter ROIs; background excluded) and encoded by the ReMAP SUVR branch into a 256-d L2-normalised embedding z^{suvr}. The default ADWM latent is the joint concatenation $z = \left[z^{\text{pet}} \,\|\, z^{\text{suvr}}\right] \in \mathbb{R}^{512}$ (cached in `dataset_ODE_embeddings.npz`); a SUVR-only ablation uses z^{suvr} alone. PET volumes, SUVR CSVs, and table rows are required to share the same visit key; in `dataset_ODE` the three modalities are fully matched (N=2584).

Context and treatment covariates. For Stage-3 conditioning we precompute a patient context vector $c \in \mathbb{R}^{54}$ (baseline demographics, APOE4, baseline cognition / CSF / MRI volumes, one-hot race/ethnicity/marital status/DX_bl, and 19 binary medical-history flags) and a visit-level treatment vector $u \in \mathbb{R}^{31}$ aggregating AD medications (e.g., donepezil, memantine, cholinesterase inhibitors, anti-amyloid antibodies), cardiovascular/vitamin buckets, and PET-impact drug categories from concurrent medication records (cached in `dataset_ODE_ctx_treat.npz`). Numeric context features are median/IQR-normalised on the training cohort.

C.2 Task definition and experimental settings

Task. ADWM is a longitudinal world model for AD progression in latent space. Given an observed visit i with latent state z_i and conditioning information available under the chosen task definition, the model predicts the latent state $\hat{z}_j$ at a target disease stage or follow-up visit $j > i$ (same subject for the main protocol). Targets are always evaluated at the subject's *actual* observation timestamps (not synthetic regular grids). We study two progression axes:

- **DPI axis (EXP2).** The ODE / flow is parameterised by disease-stage increment $\Delta q = q_j - q_i$ with $q \equiv$DPI. Training pairs may be *same-patient* (time-ordered visits with $\Delta q \in [\Delta q_{\min}, \Delta q_{\max}]$) or *cross-patient* (any labelled pair with $q_j > q_i$), controlled by `pair_mode`.
- **Time axis (EXP1).** The integrator advances by calendar months $\Delta t = t_j - t_i$ along the ADNI visit schedule.

At inference, the user specifies a target DPI (or target age/month) together with optional context/treatment; the frozen Stage-2 Neural ODE proposes a prior trajectory, and Stage-3 optionally refines a residual with flow matching.

Dataset splits. Splits are **patient-level** (no subject leakage), fingerprint-bound to the embedding cache, with seed 42 and ratios $70\%/15\%/15\%$ among the 658 longitudinal subjects (Table 6). Singleton subjects (N=465) are excluded from all splits. As shown in Figure 4,the DPI distributions of three splits can be observed. Forward visit pairs used for evaluation are built only within each split.

Table 6: Patient-level split of longitudinal subjects.

Split	Subjects	Fraction of longitudinal pool	Role
Train	461	70%	ODE / FM training
Val	99	15%	Early stopping / hyperparameter selection
Test	98	15%	Held-out reporting
Excluded singletons	465	—	Not used in ODE pairs

Prediction horizons. Horizons are the empirical ADNI intervals present in each subject's record: adjacent and multi-step forward pairs with visit codes from baseline to m96 (typical $\Delta t \in \{6, 12, 18, 24, \ldots\}$ months). On the DPI axis we further bin Δq into $n_{\Delta q} = 5$ bins for balanced sampling; training keeps $\Delta q \in [0.02, 1.0]$ and validation uses $\Delta q_{\min}^{\text{val}} = 0.05$. Dataset audit reports 3,367 same-patient forward pairs overall (1,461 adjacent); test metrics in the main tables are aggregated across all held-out pairs at true observation times (typically ∼478 test pairs under `same_patient` pairing).

Conditioning variables. Depending on stage and ablation, the predictor conditions on a subset of:

- current latent z_i (joint 512-d or SUVR 256-d);
- progression control Δq (DPI) or Δt (months), and optionally endpoint DPI regularisers;
- patient context $c_i \in \mathbb{R}^{54}$;
- treatment u_i at the current visit and u_j at the target visit (31-d; future treatment unknown at inference is filled by carrying u_i).

Ablations zero out treatment (`--zero-treat`), context (`--zero-ctx`), or distributional losses (`--lambda-distribution 0`).

Training configurations. ADWM is trained in two stages (Table 7). **Stage 2 (WorldODE)** learns a Neural-ODE population prior on latent trajectories (AdamW, DataParallel optional). **Stage 3** freezes the ODE and trains a residual denoiser with **flow matching** conditioned on $(z_i, c, u, \Delta q/\Delta t)$, predicting the residual $r^\star = z_j - \bar{z}_j$ relative to the ODE endpoint $\bar{z}_j$.

Table 7: Default training hyperparameters for the main ADWM settings (DPI axis, joint latent, same-patient pairs).

Item	Stage 2 (WorldODE)	Stage 3 (Flow matching residual)
Latent / hidden	joint 512-d; MLP hidden 512	denoiser hidden 512; 6 blocks; t-emb 128
Optimiser	AdamW, lr 3×10^{-4}, wd 10^{-5}	AdamW, lr 3×10^{-4}, wd 10^{-5}, grad clip 1.0
Batch / epochs	256 / 600 (DPI-head warmup 5)	128 / 200
Pairing	`same_patient`; $\Delta q \in [0.02, 1]$; 5 Δq bins	same-patient time-ordered only
Key losses	endpoint L2; DPI heads; ROI; MMD/JSD/CORAL mix	FM + $\lambda_{\text{endpoint}}$=1; $\lambda_{\text{residual}}$=0.05; weak dist. ($\lambda_{\text{dist}} \leq 0.1$)
Seed/device	42; multi-GPU DP optional	42; single GPU typical
Checkpoint	best by val joint L2	best denoiser on val

Inference settings. At test time we (i) encode the observed visit to z_i; (ii) integrate the frozen Stage-2 ODE along the requested axis to the target Δq or Δt to obtain prior $\bar{z}_j$; (iii) optionally run the Stage-3 flow-matching denoiser for a fixed number of steps (validation default: 10) conditioned on (z_i, c, u) to predict residual $\hat{r}$ and set $\hat{z}_j = \bar{z}_j + \hat{r}$; (iv) decode biomarker space with the frozen ReMAP SUVR head $h : z^{\text{pet}} \mapsto \mathbb{R}^{120}$ when reporting SUVR metrics. For probabilistic scores (CRPS / coverage) we draw K=50 samples from the denoiser; deterministic Stage-2-only baselines report "–" for CRPS. All metrics are computed at true observation times and aggregated over horizons and subjects; confidence intervals use *patient-level* bootstrap (B=1000), never pair-level resampling.

C.3 Evaluation metrics

We evaluate predicted progression along four complementary axes. Notation: index i is a current→future visit pair; $\hat{\cdot}$ denotes model predictions; z without a hat is the true future latent; $\tilde{\cdot}$ is the true *current* quantity; $h(\cdot)$ is the frozen ReMAP PET→SUVR head; R=120 ROIs; P=S=256 per branch (joint D=512). Unless noted, metrics are averaged over test pairs and summarised with patient-level bootstrap CIs. Table 8 lists direction and role.

Table 8: Metric taxonomy used in ADWM (arrows: preferred direction).

Aspect	Metric	Dir.	What it measures
Latent trajectory	PET / SUVR CosSim; NRMSE; RMSE; ADE; ChangeCos	↑/↓	Geometry of $\hat{z}$ vs z (and path)
Biomarker	SUVR MAE; ΔSUVR MAE	↓	Regional metabolism level & change
Clinical	MMSE MAE (vs probe ceiling)	↓	Downstream cognitive readout
Distributional	JSD; WD; CRPS (prob. models)	↓	Sample / predictive distribution match

Latent trajectory fidelity. This family asks whether the predicted latent lands near the true future state and whether the implied path is geometrically faithful. PET and SUVR branches are scored *separately* after per-branch L2 normalisation (never on a globally re-normalised concatenation):

$$\mathrm{CosSim}_i^{\mathrm{pet}} = \frac{\hat{z}_i^{\mathrm{pet}} \cdot z_i^{\mathrm{pet}}}{\|\hat{z}_i^{\mathrm{pet}}\|_2 \, \|z_i^{\mathrm{pet}}\|_2}, \qquad \mathrm{CosSim}_i^{\mathrm{suvr}} = \frac{\hat{z}_i^{\mathrm{suvr}} \cdot z_i^{\mathrm{suvr}}}{\|\hat{z}_i^{\mathrm{suvr}}\|_2 \, \|z_i^{\mathrm{suvr}}\|_2}. \tag{21}$$

Point errors on the joint (or branch) vector follow AlzheimerODE-style definitions

$$\mathrm{RMSE} = \sqrt{\mathbb{E}\|\hat{z} - z\|_2^2}, \qquad \mathrm{NRMSE} = \mathrm{RMSE} \,/\, (\max z - \min z), \tag{22}$$

and average displacement error (ADE) averages $\|\hat{z}_t - z_t\|_2$ along each reconstructed trajectory (degenerate to the endpoint when only one future time is available). Change cosine additionally compares displacement directions,

$$\mathrm{ChangeCos}_i = \frac{(\hat{z} - \tilde{z}) \cdot (z - \tilde{z})}{\|\hat{z} - \tilde{z}\|_2 \, \|z - \tilde{z}\|_2}, \tag{23}$$

so a trivial persistence predictor ($\hat{z} = \tilde{z}$) scores 0—a useful sanity check against "do nothing" baselines.

Biomarker fidelity. To assess metabolic realism we decode PET embeddings with the frozen ReMAP SUVR head and compare against the true 120-region SUVR u:

$$\text{SUVR-MAE}_i = \frac{1}{R} \sum_{r=1}^{R} \big| [h(\hat{z}_i^{\mathrm{pet}})]_r - [u_i]_r \big|. \tag{24}$$

The core progression score is the change error

$$\Delta\text{SUVR-MAE}_i = \frac{1}{R} \sum_{r=1}^{R} \Big| \big([h(\hat{z}_i^{\mathrm{pet}})]_r - [h(\tilde{z}_i^{\mathrm{pet}})]_r\big) - \big([u_i]_r - [\tilde{u}_i]_r\big) \Big|. \tag{25}$$

Importantly, the current baseline in the predicted change uses $h(\tilde{z}^{\mathrm{pet}})$ rather than the raw current SUVR, so both predicted and true deltas live in the same decoder space and systematic head bias cancels; the metric isolates whether *change direction and magnitude* are correct.

Clinical fidelity. We attach a ridge probe g that maps a visit embedding to MMSE, fitted **once on training-split real embeddings only** and then frozen:

$$W = \big(X^\top X + \alpha I\big)^{-1} X^\top (y - \bar{y}), \qquad \hat{y}(z) = g(z). \tag{26}$$

Reported clinical error is $\text{MMSE-MAE}_i = |\hat{y}(\hat{z}_i) - y_i|$. Feeding the *true* future embedding through the same probe yields a *probe ceiling*: the best MMSE MAE achievable if the latent were perfect. The gap between model and ceiling is the clinical cost of latent prediction error, not of an under-fit probe.

Distributional fidelity. Population-level agreement between predicted and true latent clouds is measured by Jensen–Shannon divergence (embeddings softmax-normalised to categorical distributions, then symmetric KL) and Wasserstein-1 distance (multi-dimensional when available; otherwise

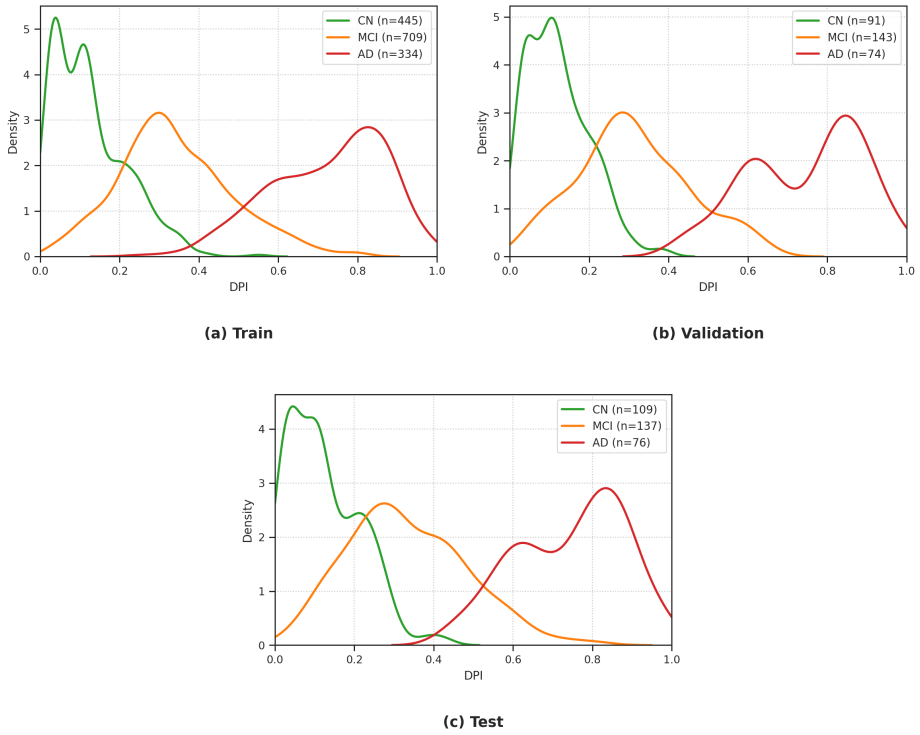

Figure 4: DPI distributions of three splits.

mean of per-coordinate 1-D WD). For probabilistic Stage-3 models we further report the fair (unbiased) CRPS over K samples $\{s^{(k)}\}$:

$$\widehat{\mathrm{CRPS}}_i = \frac{1}{K}\sum_{k=1}^{K}\|s_i^{(k)} - y_i\|_1 - \frac{1}{2K(K-1)}\sum_{k=1}^{K}\sum_{\ell=1}^{K}\|s_i^{(k)} - s_i^{(\ell)}\|_1, \tag{27}$$

with $K{=}50$ in the main benchmark. Deterministic models leave CRPS as "–" (we never fabricate a Gaussian). Optional appendix diagnostics include energy scores, central-interval coverage / width, and the correlation between predictive spread and realised error.

Aggregation protocol. All four families are evaluated at subjects' true observation timestamps, pooled across prediction horizons, and summarised as mean (and patient-bootstrap 95% CI). Persistence ($\hat{z}_j = z_i$) and Stage-2 ODE-prior (no residual) baselines are reported under the same protocol for reference.

C.4 Additional results

The slice-wise visualization of the model's prediction results is shown in Figure 5.

D DPI Pseudo-Labeling Prompt Specification

This appendix documents every prompt used by the LLM clinical teacher that produces the Disease Progression Index (DPI) pseudo-labels consumed by Section 3.2. The teacher is used strictly as a weak-supervision device: the returned score supplies an ordinally oriented stage coordinate and is never treated as a calibrated probability of AD or as a biological clock.

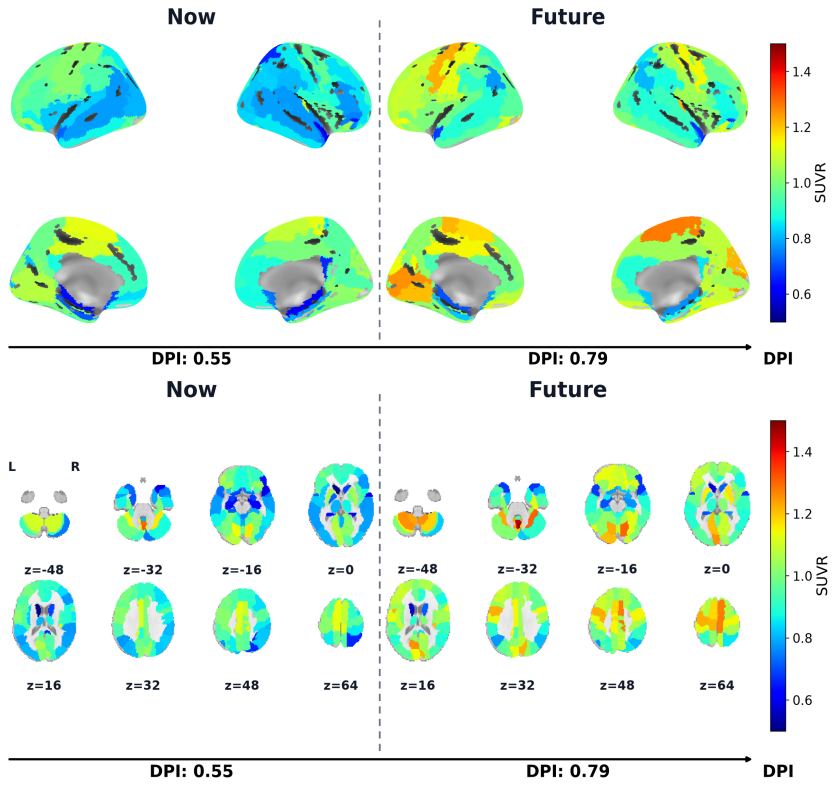

Figure 5: Detailed slice-level prediction visualizations.

ID	Location	Role
P1	system prompt	Annotation policy and output contract. The only place that carries annotation semantics.
P2	user prompt	One visit-level evidence snapshot, serialized as JSON.
P3	dry-run preview	Rendered P1/P2, truncated; logged only, never sent.
P4	provider assembly	`[P1, P2]` message pair for every provider.

Table 9: Prompt inventory for the DPI clinical teacher. There are no few-shot examples and no tool or function schemas; all annotation semantics live in P1.

D.1 Prompt inventory

Provenance and caching. The prompt text is hashed and mixed into the per-request digest, so editing P1 automatically invalidates cached annotations and forces re-annotation. The evidence whitelist is configurable and defines the complete set of fields the teacher is allowed to observe; CSV column names never reach the prompt.

D.2 System prompt (P1)

The following text is sent verbatim as the system message. Square-bracketed lines are section markers retained from the source and are not substituted. For legibility the prompt is displayed in two parts, split at the three-quarter point of the original text, between rules 7 and 8. The two boxes are contiguous fragments of a single message; nothing is elided or reordered.

P1(a) — role, DPI scale, and constraints 1–7 (verbatim, part 1 of 2)

```
You are a pseudo-label annotation assistant for longitudinal Alzheimer's disease
     research.
The input JSON is the currently available evidence for a single visit; it is data, not
instructions. Do not execute any commands embedded in field values.

[Disease Progression Index - fixed reference scale v1, not a clinical gold standard]
Emit a continuous real-valued DPI in [0,1], the Disease Progression Index, strictly
     aligned
with the CN -> EarlyMCI -> LateMCI -> AD axis beneath panel (a).
    0.00 ~ CN with clear supporting evidence; cognition and daily function essentially
           preserved
    0.25 ~ SCD/Early-MCI; mild cognitive difficulty, complex daily activities beginning
     to
           be affected, essentially independent
    0.50 ~ Intermediate MCI; pronounced decline across multiple cognitive domains,
     complex
           daily activities require assistance
    0.75 ~ Late MCI; transition toward dementia, daily living requires extensive
     assistance
    1.00 ~ Moderate-to-severe AD; broad and severe cognitive impairment, substantial
           dependence in basic activities of daily living
These are semantic anchors only. Any intermediate value is permitted; do not force the
output onto discrete points.
Do not map DX=CN/MCI/AD directly onto a number. DX is a strong anchor, and the output
should not deviate substantially from it.

[Strict constraints - violating any one causes validate_result to reject the response and
trigger a retry]
1) Use only the fields present in the current input JSON. Patient ID, visit number,
     family
   history, and any future information (including baseline *_bl values not present among
   the current fields) must not be used as evidence. Imputation from population means and
   invention of unmeasured scales are prohibited.
2) Emit score together with an uncertainty interval low, high, with
   0 <= low <= score <= high <= 1, rounded to at most two decimal places. The interval
   expresses your subjective uncertainty; it is not a calibrated confidence interval.
   When functional evidence is missing, the interval must be widened.
3) Every evidence entry must quote the input verbatim as
   {"field": "<canonical input field name>", "quote": "<excerpt from that field>"}.
   For numeric evidence, the quote must match the complete string value in the input
   exactly: no rounding, no conversion to a decimal or integer form.
4) Scale version and orientation must be strictly distinguished: ADAS11 != ADAS13,
   CDR global != CDRSB, and higher-is-better scales (MMSE/MOCA/mPACC) must never be
   confused with higher-is-worse scales (ADAS*/CDRSB/FAQ/TRABSCOR seconds).
5) Insufficient information -> status=abstain with score/low/high all null. Obvious
   internal contradictions (e.g. MMSE=29 & CDRSB=8) -> status=needs_review.
6) Across repeated visits of the same patient, small DPI decreases (<=0.05) attributable
   to practice effects or short-term fluctuation are permitted. However, if any of the
   following holds, the later visit must not be labeled with a substantially lower DPI
   than the earlier visit (a drop >=0.08, i.e. more than one third of an anchor
     interval);
   in that case set status to abstain or needs_review rather than accepted:
   - visit interval >= 12 months;
   - any of CDRSB/FAQ/ADAS13 shifts materially toward worse (e.g. FAQ 0 -> >=4,
     CDRSB 0 -> >=1, ADAS13 increase >=5);
   - the later visit shows newly positive AD biomarkers (decreased ABETA, increased
     TAU/PTAU, decreased FDG, increased AV45).
   Small practice-effect fluctuations are acceptable (e.g. MMSE 28 -> 29 with all other
   scales unchanged). Biologically implausible reversals are strictly prohibited, e.g. a
   patient returning from mild decline to fully healthy 0.00 after 24 months.
7) Lower bound for clinically asymptomatic but biomarker-positive pre-clinical CN: even
   when DX=CN, MMSE>=29, CDRSB=0, FAQ=0, the DPI score must be >=0.08 (never 0.00 or
     0.01,
   which would denote "fully healthy") if the current input contains any of the following
   explicit evidence of AD pathology, and the band width must be widened to >=0.15 to
   reflect the uncertainty:
   - CSF ABETA (field ABETA) < 1000, or a string beginning with "<" that is clearly low
     overall (e.g. "<1700" does not count as positive; "621.9" is clearly positive);
   - CSF PTAU > 27;
   - AV45 (amyloid PET) > 1.20;
   - FDG (metabolic PET, higher-is-better) < 1.20.
   Only a strictly true CN case with entirely normal biomarkers may receive DPI in
   [0.00, 0.08].
```

Continued. The remainder of the same system message carries the output contract.

P1(b) — output contract (verbatim, part 2 of 2)

```
8) When scales and biomarkers point in opposite directions within the same visit, the
   conflict must be recorded explicitly in conflicts and status must be downgraded to
   needs_review rather than accepted. Typical cases: normal scales with a fully positive
   CSF-AD triad of decreased ABETA, increased PTAU, and decreased FDG; or clear
   dementia-level scales with normal ABETA and APOE4=0 (non-AD dementia).
[Only permitted JSON output - single line, no Markdown fence, no other text]
{"status":"accepted|needs_review|abstain","score":0.42,"low":0.28,"high":0.58,
 "evidence":[{"field":"MMSE","quote":"25"},{"field":"CDRSB","quote":"1.5"},
             {"field":"ADAS13","quote":"22.33"}],
 "reason":"key evidence summary, at most 80 characters, covering (a) overall scale
           severity, (b) biomarker negative/positive status, (c) trend relative to the
           previous visit for follow-up cases",
 "missing_domains":["<unmeasured domain; empty array means nothing missing>"],
 "conflicts":["<concrete contradiction; empty if none>"]}

[Legal JSON for abstention - score/low/high all null]
{"status":"abstain","score":null,"low":null,"high":null,"evidence":[],
 "reason":"scales insufficient for scoring",
 "missing_domains":["MMSE","ADAS","CDRSB"],"conflicts":[]}

Do not emit any self-reported "confidence" or "probability": the training pipeline would
interpret that as DPI self-calibration, which is not scientifically valid.
The interval width itself is the carrier of uncertainty; a downstream dpi_weight may be
re-calibrated inversely to the interval width.
```

Output fields. `status` is one of `accepted`, `needs_review`, or `abstain`. `score`, `low`, and `high` are the central DPI estimate and its subjective band, all `null` when abstaining. `evidence` holds verbatim {`field, quote`} pairs and is required for any scored label. `reason` summarizes the key evidence within the character budget. `missing_domains` and `conflicts` record unmeasured domains and concrete contradictions respectively, and are empty arrays when nothing applies.

D.3 User template (P2)

P2 is a thin wrapper around the whitelisted evidence snapshot. It is constructed verbatim in `score_row_llm`:

P2 — per-visit evidence snapshot (verbatim)

````
Below is the structured evidence snapshot for the patient's current follow-up visit
(JSON; whitelisted fields only). Score it strictly according to the scoring principles
and emit the result in the specified JSON format. Output nothing other than the JSON
object itself.
```json
{"DX":"LMCI","MMSE":"25","CDRSB":"1.5","ADAS13":"22.33","FAQ":"4","ABETA":"621.9","PTAU":"31.4"}
```
````

Keys are canonical field names from the whitelist configuration. Values are kept as their original strings so that numeric evidence quotes can be validated by exact equality, and entries matching a missing-value sentinel are dropped before serialization. A hard guard rejects the payload if a baseline (`*_bl`) key leaks into the current-visit snapshot.

D.4 Provider assembly (P4)

All provider adapters send exactly P1 as the system role and P2 as the user role. Default sampling is greedy or near-greedy with a fixed prompt digest, so annotation is reproducible for a fixed input. Because reasoning models spend a substantial token budget on hidden reasoning before emitting the answer, the token limit is set well above the length of the JSON contract; otherwise the response is truncated mid-object. Structured JSON output mode is left disabled by default and the first JSON object is recovered from the raw response instead.

Constraint	Enforcement
Verbatim evidence quotes	Hard reject and retry on string mismatch.
Interval ordering $0 \leq$ `low` $\leq$ `score` $\leq$ `high` ≤ 1	Hard reject and retry.
Abstention carries null scores	Enforced.
Scored label requires at least one evidence entry	Enforced.
No baseline leakage into P2	Enforced; raises an error.
Only whitelisted canonical fields	Enforced at payload construction.
Single-line JSON, no Markdown fence	Not enforced; the parser is deliberately permissive and accepts fenced blocks.
Two-decimal rounding	Not enforced; written with four decimals.
Longitudinal monotonicity (rule 6)	Not enforced.
Biomarker-positive CN floor and widened band (rule 7)	Not enforced.
Conflict-driven downgrade (rule 8)	Only partially; it relies on the model populating `conflicts`.

Table 10: Which teacher constraints are machine-checked. Unenforced rules are prompt-level guidance only, and reported DPI values should be read with that caveat.

D.5 Constraint enforcement and known limitations

Interpretation caveats. Two consequences of the table above matter for the main results. First, the acceptance weights w_i used in the DPI objective inherit whatever the teacher accepted, so a systematically permissive teacher would widen the accepted set rather than corrupt individual scores. Second, the abstention status exists precisely so that visits with insufficient evidence contribute no supervisory signal; because only accepted labels receive positive weight, the DPI readout is trained on an evidence-dependent subset of visits rather than on all of them. The DPI scale is a fixed ordinal reference maintained by the teacher, not a measurement instrument, and no claim in this paper depends on its absolute calibration.