EVOLVING ANATOMY GRAPHS STEER LONGITUDINAL IMAGE FORECASTING BEYOND PIXELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Longitudinal image forecasting aims to predict how a patient's anatomy will change, yet image similarity can reward a forecast that changes nothing. On our brain MRI benchmark, copying the baseline outperforms every tested forecaster on PSNR and SSIM despite predicting no progression. The forecasters also miss the change itself: on three core brain structures, none reaches the change correlation of a regression on elapsed time (0.324). We introduce **GraphMorph**, which makes regional geometry an explicit forecast. Each anatomical region is encoded as a spatial node in an anatomy graph, retaining shape and location information while allowing its evolution to depend on other regions. Conditioned on the requested interval, clinical covariates and available anatomical history, the model evolves these nodes under supervision from measured follow-up anatomy and decodes them into spatial maps of the predicted future anatomy, the *anatomical request*. A separately trained renderer deforms the patient's baseline scan using the request as both spatial control and its primary supervision target. The anatomical forecast therefore determines not only what the renderer receives, but what it learns to reproduce, allowing forecast accuracy and rendering fidelity to be assessed separately. On brain MRI and knee radiographs, GraphMorph achieves regional-change correlation 0.314 over 32 brain regions, 0.422 on the core structures and 0.230 for knee joint space, versus 0.230, 0.268 and 0.119 for the strongest of ten brain and six knee image forecasters. Substituting measured follow-up anatomy into the same frozen renderer raises core-brain change correlation from 0.422 to 0.704, showing that the generated change responds to the supplied anatomy.

1 INTRODUCTION

Managing degenerative disease requires understanding both a patient's current condition and how it is likely to change. Serial imaging tracks anatomical progression, including brain atrophy in Alzheimer's disease and joint-space narrowing in osteoarthritis (Frisoni et al., 2010; Hunter & Bierma-Zeinstra, 2019). Anticipating these changes could inform closer monitoring and help researchers identify trial participants likely to progress during the study interval (Puglisi et al., 2025). Longitudinal image forecasting seeks to predict this progression in the form of a future scan: the image must express the patient's expected change, not simply resemble a plausible follow-up.

Learning progression from serial images is difficult because differences between visits reflect disease-related anatomical changes as well as acquisition conditions and residual misalignment. The anatomical changes are small: in our brain cohort, the hippocampus, amygdala, and lateral ventricles change by approximately 2–6% of their own volume between visits, with a median interval of seven months. A whole-image reconstruction objective combines errors in these regional changes with differences elsewhere in the image. It does not explicitly distinguish the anatomical progression to be predicted from other inter-visit variation.

Recent methods generate future scans using latent diffusion, anatomical conditioning, or flow matching (Puglisi et al., 2025; Litrico et al., 2024; Chen et al., 2026). In these models anatomy enters as a conditioning signal: BrLP conditions generation on disease-trajectory estimates, and Mamba-Control incorporates a graph of image-patch features into its generator (Puglisi et al., 2025; Yang et al., 2025). A complementary line deforms baseline scans to realise prescribed anatomical changes

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.

(Khanal et al., 2017; Pombo et al., 2023). None of these methods forecasts future regional geometry as a separate, supervised output that the generator is trained to reproduce. GraphMorph does both, linking anatomical forecasting with prescribed-anatomy rendering.

We use regional geometry to specify progression before generating an image. **GraphMorph** separates predicting future anatomy from rendering it. The forecaster is trained to predict future geometry from baseline anatomy, a requested interval and baseline clinical information, supervised by follow-up segmentations. Earlier visits supply history when available. The prediction guides a deformation of the baseline scan, which supplies image content.

Regional volume forecasts specify size but leave future boundary locations undetermined. We represent each region as a spatial node in an *anatomy graph*, retaining shape and location while allowing its temporal update to depend on other regions. A residual evolver advances the states jointly over the requested interval, starting from persistence and learning departures from it. The states decode into spatial maps of future anatomy, the *anatomical request*; differences from decoded baseline maps describe progression. Supervision on future-anatomy maps trains the evolver directly on regional geometry. The frozen renderer then receives the request through a dedicated conditioning branch, which guides the future scan generation to match the requested anatomy.

To summarize, our contributions are:

1. **Learning progression as spatial anatomical evolution.** GraphMorph represents regional anatomy as spatial graph states and learns their temporal evolution under direct anatomical supervision. The decoded forecast specifies future regional geometry, rather than only scalar volume changes or high-dimensional latent features, making the predicted progression spatially explicit.
2. **Prescribed-anatomy rendering with a frozen deformation renderer.** We pretrain a deformation renderer for longitudinal registration and then freeze it. The predicted anatomical request is introduced through a separate conditioning branch that steers the frozen renderer toward the requested geometry. This interface allows the anatomical forecast to be evaluated and replaced while the renderer remains fixed.
3. **Anatomy-level gains that reach the image.** On brain MRI and knee radiographs, GraphMorph achieves regional-change correlation 0.314 over 32 brain regions and 0.230 for knee joint space, versus 0.230 and 0.119 for the strongest image forecasters (Section 5; Appendix E). Controlled substitutions show that the same frozen renderer recovers more of the measured change when supplied with more accurate anatomy, and request quality tracks rendered-change accuracy across the tested forecasters (Figure 4).

2 Related Work

Longitudinal image forecasting. TADM predicts inter-visit image residuals (Litrico et al., 2024), ImageFlowNet models image trajectories with neural ODEs (Liu et al., 2025), and Δ-LFM learns changes between visit latents through flow matching (Chen et al., 2026). BrLP conditions latent diffusion on disease-trajectory estimates (Puglisi et al., 2025). GraphMorph predicts regional geometry before image generation, with anatomical evolution supervised separately from image reconstruction.

Anatomical guidance and supervision. DANI-Net constrains generated progression with regional atrophy rates (Ravì et al., 2022), while AG-LDM uses a tissue-segmentation loss to supervise diffusion forecasting (Wan et al., 2026). MambaControl uses a Fourier-refined graph of image-patch features to condition its generator (Yang et al., 2025). PCGM uses a structural causal graph of metadata and scalar regional volumes; metadata interventions determine volumes, which are converted into masks by boundary editing to guide diffusion (Li et al., 2025). Both methods condition generation on explicit anatomy. In GraphMorph, each neural graph node encodes a region's segmentation as a spatial tensor, and cross-region attention guides its evolution over the requested inter-visit interval. The decoded future anatomy is learned from longitudinal segmentations and serves as the renderer's primary anatomical training target. Appendix K.2 details the differences in graph computation, geometry prediction, supervision and evaluation.

Rendering prescribed anatomy. Atrophy simulators and CounterSynth deform a subject's scan to realise prescribed anatomical changes (Khanal et al., 2017; Pombo et al., 2023). GraphMorph retains this separation between specifying anatomy and rendering, but learns future geometry from longitudinal observations. We reuse DiffuseMorph's pretrained diffusion and deformation networks (Kim et al., 2022), replacing target-image conditioning with a predicted anatomical request. Generation is trained primarily to reproduce that request; the follow-up image is unavailable at inference. We follow change-sensitive evaluation alongside image fidelity (Feng et al., 2026; Disch et al., 2025; Farki et al., 2026). Broader progression models and forecasting literature are reviewed in Appendix K.

3 METHODOLOGY

GraphMorph predicts future anatomy through spatial graph evolution and uses the decoded prediction as both a conditioning input and the primary anatomical supervision target for image generation (Figure 1). We train the anatomical representation, temporal evolver, and renderer sequentially, keeping previously trained components fixed. Follow-up segmentations supervise the forecaster; its predictions supervise the renderer.

Forecasting task. Let M be a baseline scan, F its follow-up, and Δa the elapsed interval. The available inputs are M, its segmentation S_M, baseline clinical covariates b, and Δa. When an earlier visit is available, q contains pooled latent velocity and regional change rates over the preceding interval. The outputs are spatial maps of the predicted future anatomy and a generated scan $\hat{F}$. The follow-up F and its segmentation S_F supply training targets, never inference inputs. We describe the brain instantiation with $K = 32$ SynthSeg regions (Billot et al., 2023); the knee instantiation is detailed in Appendix J.2. The hippocampus, amygdala, and lateral ventricles form the core evaluation structures, with the temporal horn reported separately.

3.1 SPATIAL ANATOMY GRAPHS AND TEMPORAL EVOLUTION

Spatial regional representation. Each region's binary mask is encoded into a spatial latent. Collecting the K nodes gives $Z = E(S) \in \mathbb{R}^{K\times c\times d\times h\times w}$, with c channels. A graph encoder G exchanges information between nodes and returns $H = G(Z)$ of the same shape. The decoder D maps each node back to regional occupancy logits at image resolution. Applying a sigmoid gives a soft map of the region's spatial extent. These maps describe anatomy; their differences between visits describe anatomical change.

The spatial grid lets a node encode boundary shape and location rather than only regional size. Graph interactions provide cross-region context without removing that spatial state. Stage 1 trains E, G, D on individual segmentations using soft Dice, binary cross-entropy, and an absolute log-volume-ratio loss. The overlap terms constrain geometry, while the volume term explicitly penalises regional size error. The architecture, loss weights, and reconstruction analysis are given in Appendices I and G.1.

Graph-conditioned temporal evolution. With E, G, D frozen, the evolver predicts $\hat{H}_F = \mathcal{P}(H_M, \Delta a, b, q)$ from $H_M = G(E(S_M))$. At step j, each spatial state $H_k^{(j)}$ is pooled and embedded with its region identity and available history into a summary $s_k^{(j)}$. Clinical conditioning supplies the input-visit diagnosis and age advanced to the current step (Appendix I.7). For one attention head, the weight assigned by region k to region ℓ has the form

$$\alpha_{k\ell}^{(j)} = \mathrm{softmax}_\ell\left(\frac{\langle W_Q \tilde{s}_k^{(j)}, W_K \tilde{s}_\ell^{(j)}\rangle}{\sqrt{d_a}} + B_{k\ell}\right), \tag{1}$$

where $\tilde{s}$ is the clinically and temporally conditioned summary, d_a is the head dimension, and B is a learned region-pair bias shared across heads. Attention aggregates value projections across regions; a linear readout after the attention blocks produces channel-wise coefficients $\gamma_k^{(j)}, \beta_k^{(j)}$. These modulate the spatial update:

$$H_k^{(j+1)} = H_k^{(j)} + \Delta t\, \mathrm{Conv}\Big(H_k^{(j)} \odot (1 + \gamma_k^{(j)}) + \beta_k^{(j)}\Big), \tag{2}$$

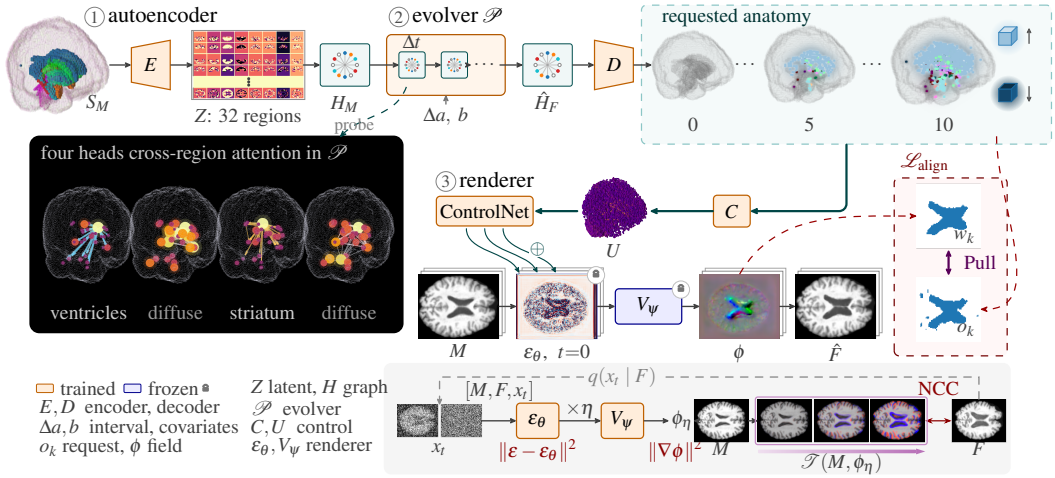

Figure 1: GraphMorph: spatial anatomical encoding, temporal evolution and request-guided rendering. The training band shows registration pretraining; image panels show one test pair and the dark panel an offline attention probe (Appendix I.3).

with $H^{(0)} = H_M$, $H^{(n)} = \hat{H}_F$ and $\Delta t = \Delta a/n$. The coefficients are broadcast over each node's spatial grid. Cross-region summaries modulate the update while each node retains its evolving spatial grid. The increment convolution is zero-initialised, so the initial forecast is persistence. History features, attention blocks and step selection are specified in Appendix I.

Supervising the forecast. The objective combines overlap and log-volume losses against S_F, an auxiliary SmoothL1 loss against $G(E(S_F))$, and a cross-pair correlation loss on decoded regional change. Overlap and volume terms constrain individual predictions; correlation targets variation across pairs. Loss weights, region sampling and correlation estimation are in Appendix I.

3.2 Anatomical Requests for Image Generation

Future anatomy and progression maps. The frozen decoder maps the baseline and predicted states into $O_0 = \mathrm{sigm}(D(H_M))$ and $O_{\Delta a} = \mathrm{sigm}(D(\hat{H}_F))$. Each has one channel per region at the resolution of M. We call $O_{\Delta a}$ the *anatomical request*: it specifies predicted future geometry. The difference $O_{\Delta a} - O_0$ gives the *per-region progression maps*, indicating where a region is predicted to gain or lose occupancy. A learned $1 \times 1 \times 1$ convolution C builds the control field:

$$U = C\big([\,O_0,\ O_{\Delta a} - O_0\,]\big). \tag{3}$$

The renderer receives current anatomy and the requested change in image coordinates; Table 4 compares this decoded control with latent control.

Rendering the request. The renderer reuses a DiffuseMorph score network ϵ_θ and deformation network V_ψ, pretrained for registration on training-split pairs (Kim et al., 2022). A ControlNet-style branch injects U into the score network's skip connections (Zhang et al., 2023). The pretrained networks and the forecaster remain frozen; only C and the control branch train. Rendering uses a deterministic evaluation of the score network at diffusion time $t = 0$ on $[M, M, M]$, not iterative diffusion sampling. This replaces the follow-up channel used during registration with the baseline (Appendix I.4). The deformation network produces a field ϕ, and a spatial transformer $\mathcal{T}$ gives $\hat{F} = \mathcal{T}(M, \phi)$. The baseline supplies image content; U is the only input carrying the forecast.

Anatomical supervision. Let Y_M and Y_F be the per-region indicator maps of S_M and S_F. The principal anatomical objectives of forecasting and rendering are

$$\mathcal{L}^{\text{anat}}_{\text{forecast}} = \ell_f(O_{\Delta a}, Y_F), \tag{4}$$

$$\mathcal{L}^{\text{anat}}_{\text{render}} = \ell_r(\mathcal{T}(Y_M, \phi), \mathrm{sg}[O_{\Delta a}]), \tag{5}$$

where ℓ_f and ℓ_r combine spatial overlap and regional volume terms with stage-specific weights, and sg denotes a fixed target with no gradient to the forecaster. The expanded renderer loss is given

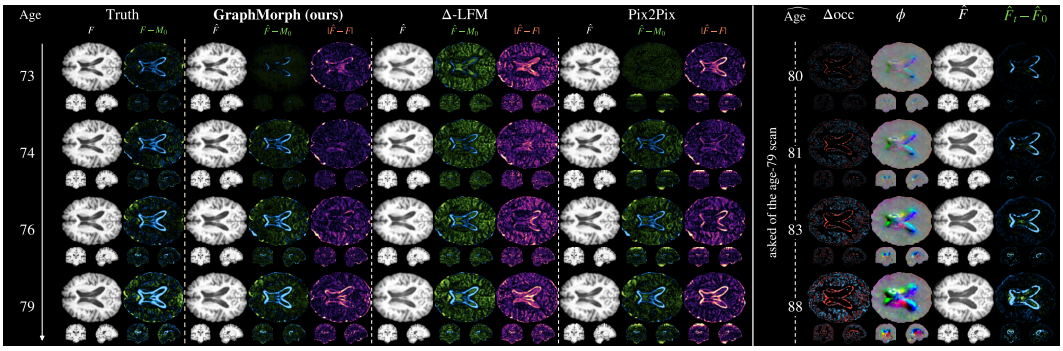

Figure 2: Method comparison and horizon response on a test subject selected for between-method core-error separation. Left: forecasts use the preceding observed visit, with cumulative changes relative to the first visit (M_0) and absolute errors against the follow-up. Right: forecasts from the age-79 scan, with changes relative to its zero-interval output. Signed changes are blue for decrease and green for increase. Views: axial (large), coronal and sagittal (small). Display scales are given in Appendix H.7.

in Eq. 10. Forecasting is trained against measured future anatomy; rendering is trained against the predicted anatomy. The rendering target is the request's own geometry, and Y_F never replaces it.

The renderer also receives deformation smoothness regularisation and a weak photometric loss against F, weighted at $1/200$ of the original registration objective's normalised-cross-correlation term. These are auxiliary losses: F is never a renderer input. The deployed renderer trains on predicted requests only; follow-up-derived training requests are examined in Appendix B.5.

3.3 Controlled anatomical substitutions

With the renderer fixed, **oracle control** supplies encoded follow-up anatomy, the **identity evolver** requests decoded baseline anatomy, and **zero control** removes the control field. Oracle control is a follow-up-informed diagnostic. These substitutions separate forecast accuracy from request adherence (definitions in Appendix I.5, scale sweep in Appendix B.3). The renderer expresses change as a deformation of the baseline.

4 Experimental Setup

Data. Two longitudinal cohorts with patient-disjoint splits: brain MRI from ADNI, OASIS-3 and AIBL (1,857 subjects; 669 test pairs) (Mueller et al., 2005; LaMontagne et al., 2019; Ellis et al., 2009) and OAI knee radiographs (8,295 knees; 7,517 strict-QC test pairs) (Nevitt et al., 2006; Peterfy et al., 2008). Cohort construction, preprocessing and quality control are in Appendix J.

Evaluation. Anatomical-change accuracy is measured by mean regional Pearson correlation (R_{chg}), change MAE in percentage points (pp), unscaled Δ_{RMAE} and retrospectively amplitude-matched Δ_{Region}. Request adherence uses $\mathrm{MAE}_{\mathrm{adh}}$, and image fidelity uses PSNR/SSIM (definitions and statistical conventions in Appendices H.2 and H.3).

Comparison protocol. No forecast uses follow-up images; follow-up-informed diagnostics are marked (F). BrLP and Δ-LFM also receive target-visit diagnosis information; GraphMorph uses the input-visit diagnosis alone (Appendix A.1). Inputs, training schedules and checkpoint rules for every method are specified in Appendices A.1, H and I.6 (Tables S1 and S36).

5 Results

5.1 Anatomical-change prediction on brain MRI and knee radiographs

Brain MRI. On the three core structures, no compared image forecaster reaches the elapsed-interval reference, 0.324 when validation-selected and 0.296 when fitted on test. GraphMorph

Table 1: Brain progression prediction. $^{***}p < 0.001$ vs. elapsed interval (test fit) (uncorrected; GraphMorph remains significant on both aggregates after Bonferroni correction over 22 tests). Test: 669 pairs from 185 subjects; core progression uses 590 pairs. ± is SD across three core structures for R_{chg}/MAE, and across pairs for progression errors and image metrics. Bold: best among comparison methods; Δ_{Region} is amplitude-matched.

	Method	Type	$R_{\mathrm{chg}}^{(\mathrm{core})}$ ↑	$\Delta_{\mathrm{RMAE}}^{(\mathrm{core})}$ ↓	$\Delta_{\mathrm{Region}}^{(\mathrm{core})}$ ↓	$\mathrm{MAE}^{(\mathrm{core})}$ ↓	PSNR↑	SSIM %↑
Comparison methods	DiffuseMorph (Kim et al., 2022)	Deform	0.077±0.077	0.927±0.702	1.094±0.628	4.00±0.64	26.33±1.86	94.50±1.98
	BrLP (Puglisi et al., 2025)	Noise	0.094±0.065	1.812±2.155	1.043±0.613	8.29±1.53	24.71±2.43	89.71±11.5
	MambaControl backbone (Yang et al., 2025)	Noise	0.003±0.076	1.237±1.038	1.311±0.560	6.14±0.85	27.07±1.81	94.58±3.22
	Δ-LFM (Chen et al., 2026)	ΔFlow	0.204±0.146	1.166±1.019	0.987±0.607	4.83±0.87	25.96±1.54	94.75±3.89
	CardiacAging (Campello et al., 2022)	Direct	0.268±0.174	0.927±0.315	1.146±0.553	3.49±1.10	27.36±2.26	95.79±2.16
	ImageFlowNet (Liu et al., 2025)	ODE	0.265±0.195	1.004±0.385	1.454±0.558	3.74±1.39	25.77±1.58	95.17±2.13
	Pix2Pix (Isola et al., 2017)	Direct	0.070±0.064	1.153±0.534	1.600±0.548	4.49±1.37	24.75±1.20	95.14±1.88
	Noise-Free Flow (Moschetto et al., 2025)	ΔFlow	0.185±0.078	0.747±0.501	0.920±0.583	3.24±0.85	27.53±2.35	95.48±2.35
	Flow Matching (Moschetto et al., 2025)	Flow	0.202±0.035	0.798±0.646	0.947±0.572	3.46±0.90	23.91±1.17	90.52±4.81
	Diffusion (Moschetto et al., 2025)	Noise	0.173±0.043	0.844±0.662	0.979±0.588	3.63±1.00	25.10±1.60	92.73±2.22
	GraphMorph (ours)	Deform	**0.422*****±0.136	**0.677**±0.447	**0.855**±0.590	**2.66**±0.73	**29.18**±2.87	**96.80**±2.12
References	*Copy-forward*	—	—	1.000±0.000	2.000±0.000	3.50±1.43	29.89±3.13	96.81±2.15
	Elapsed interval (test fit)	—	0.296±0.205	0.736±0.413	—	2.74±0.72	—	—
	Elapsed interval (validation-selected)	—	0.324±0.177	—	—	—	—	—
	Core-feature tabular	—	0.459±0.101	0.679±0.476	—	2.61±0.73	—	—
	All-region tabular	—	0.493±0.085	—	—	—	—	—
	Measurement ceiling	—	0.771	—	—	—	—	—
	Oracle control (F)	Deform	0.704±0.215	0.490±0.390	0.488±0.468	2.61±0.45	29.25±2.59	96.87±1.83

Table 2: FastSurfer remeasurement on brain test pairs. Correlation differences are GraphMorph minus Δ-LFM, and raw-MAE differences are GraphMorph minus copy-forward. Brackets: 95% paired subject-bootstrap CIs. ROI sets are given by row. Retained-subset rescoring is post hoc.

Analysis	Patients / pairs	ΔR_{chg} [95% CI]	ΔMAE (pp) [95% CI]
Compatible ROI mapping, retained subset re-scored (29 regions)	100 / 100	−0.000 [−0.062, +0.063]	+0.229 [+0.057, +0.404]
Exclude non-equivalent CSF, retained subset re-scored (31 regions)	100 / 100	+0.007 [−0.055, +0.071]	+0.235 [+0.069, +0.403]
Independent sample, pre-specified (29 regions, complete pairs)	84 / 292	+**0.042** [+0.003, +0.086]	−0.031 [−0.140, +0.080]
Independent sample, failures as copy-forward (29 regions)	85 / 296	+0.042 [+0.006, +0.085]	−0.026 [−0.135, +0.082]

reaches 0.422 on the core, a subject-bootstrap margin of +0.153 [+0.102, +0.208] over CardiacAging, the strongest of them (Table 1). Over 32 regions it reads 0.314 against 0.230 for Δ-LFM (+0.084 [+0.052, +0.133]) and 0.174 for elapsed time (Table S23), and its lead over Δ-LFM holds in every diagnosis and sex stratum (Table S22). The margin over image forecasters comes from forecasting anatomy. The all-region tabular reference predicts one scalar per region: it regresses each region's volume change from measured volumes, rates and covariates, reading 0.369 and 0.493 on the same aggregates (Appendix A.3). GraphMorph predicts how each region's spatial extent changes, of which the volume change is one summary, and renders it into the scan. On validation, before rendering, its volume forecast is within 0.015 of the tabular reference (−0.015 [−0.034, +0.010]) (Section 6). Among image forecasters GraphMorph also has the lowest core progression errors (Table 1; $\Delta_{\mathrm{RMAE}}^{(\mathrm{core})}$ 0.677, level with the core-feature tabular reference's 0.679) and the lowest 32-region volume MAE, 0.28 pp below copy-forward (Table S26). PSNR and SSIM do not register this progression: copy-forward exceeds every forecaster on both, GraphMorph coming closest, while regional volume errors span a factor of 3.5. At a median interval of about seven months the boundary change to forecast is a fraction of the 1.5 mm voxel, with the baseline labels within 0.35 mm of the follow-up labels (Appendix C.1), so a pixel-space forecaster has to recover it from voxel differences that also carry the acquisition and alignment differences that remain after registration to the common grid and visual quality checks (Appendix J).

Independent measurement. The correlation lead survives a segmenter that took no part in training. On the pre-specified independent sample of 84 patients (292 pairs), FastSurfer measurement (Henschel et al., 2020) on the compatible 29-region map gives +0.042 [95% CI +0.003, +0.086] against Δ-LFM, and the raw-MAE difference against copy-forward is −0.031 [−0.140, +0.080] (Table 2). The two segmenters agree on measured change at $r = 0.553$ on the compatible map and 0.701 on the core, and both segmenters read GraphMorph's rendered images nearly as consistently as real scans (Appendix C.1).

Knee radiographs. GraphMorph transfers to the three-region knee task after retraining: $R_{\mathrm{chg}}^{(\mathrm{JS})} = 0.230$ against 0.119 for the strongest image forecaster, with positive paired margins over all six (Tables 3 and S29), and the lowest $\Delta_{\mathrm{RMAE}}^{(\mathrm{JS})}$ and $\Delta_{\mathrm{Region}}^{(\mathrm{JS})}$ among image forecasters on the 2,850 pairs

Table 3: Knee joint-space progression. $^{**}p < 0.01$, $^{***}p < 0.001$ vs. elapsed interval (uncorrected); †: the 95% paired CI for the difference from GraphMorph excludes zero. Test: 7,517 strict-QC pairs from 835 patients. R_{chg}/MAE use stored-mask truth; progression errors use read-back truth on 2,850 pairs. The tabular reference retains the stored-mask baseline. Correlation ± is pair-bootstrap SD; other ± values are pair SD. Bold: best among comparison methods; Δ_{Region} is amplitude-matched.

	Method	Type	$R^{(\text{JS})}_{\text{chg}}$ ↑	$\Delta^{(\text{JS})}_{\text{RMAE}}$ ↓	$\Delta^{(\text{JS})}_{\text{Region}}$ ↓	MAE$^{(\text{JS})}$ (pp)↓	PSNR↑	SSIM %↑
Comparison methods	DiffuseMorph (Kim et al., 2022)	Deform	0.069±0.017	0.989†±0.532	1.337†±0.704	10.03±9.59	**22.89**±6.89	64.06±25.22
	ImageFlowNet-GraphODE (Liu et al., 2025)	ODE	0.104**±0.023	1.110†±0.861	1.262†±0.744	12.54±11.13	22.17±5.02	65.36±22.25
	Pix2Pix (Isola et al., 2017)	Direct	0.093**±0.022	1.174†±1.149	1.243†±0.731	13.53±16.25	19.20±5.34	**65.57**±26.05
	Noise-Free Flow (Moschetto et al., 2025)	ΔFlow	0.087**±0.018	0.931†±0.419	1.271†±0.716	9.63±9.13	22.16±7.26	63.60±26.98
	Flow Matching (Moschetto et al., 2025)	Flow	0.119**±0.024	1.031†±0.969	1.220±0.744	11.72±11.15	21.70±5.67	65.05±25.34
	Diffusion (Moschetto et al., 2025)	Noise	0.090**±0.022	1.072†±0.909	1.226±0.741	12.47±13.78	13.12±4.00	33.14±19.86
	GraphMorph (ours)	Deform	**0.230*****±0.016	**0.899**±0.533	**1.194**±0.741	**9.53**±8.87	21.95±6.35	63.57±25.49
References	*Copy-forward*	—	0.022±0.023	1.000†±0.005	1.998†±0.013	8.91±8.72	23.38±7.24	64.06±25.67
	Elapsed interval	—	0.038±0.011	—	—	8.92±8.70	—	—
	Tabular reference	—	0.474±0.019	0.902±0.441	1.154†±0.739	8.01±7.49	—	—
	Read-back ceiling (F)	—	0.650±0.011	—	—	6.89±6.76	—	—
	Oracle control (F)	Deform	0.755±0.010	0.562†±0.525	0.678†±0.625	5.77±6.20	23.79±7.29	65.33±25.55
	DiffuseMorph, registration (F)	Deform	0.377±0.024	0.600†±0.671	0.688†±0.629	8.85±10.23	24.69±7.87	70.62±26.30

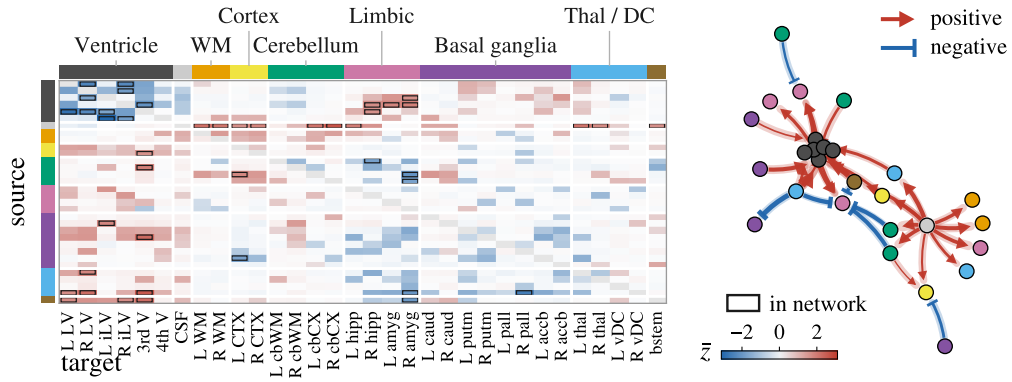

Figure 3: Fitted region-pair biases across seeds. Left: mean SD-scaled biases across six seeds; faded cells lack unanimous sign agreement. Boxed pairs form the network on the right, where edge cores and halos show the minimum and maximum bias magnitudes across seeds.

measured through the evaluation segmenter. Its unscaled joint-space error is level with the tabular reference ($\Delta^{(\text{JS})}_{\text{RMAE}}$ 0.899 against 0.902).

5.2 Accuracy of the anatomical request

The request carries the forecast. The Stage 1 autoencoder returns measured regional change at $R^{(32)}_{\text{chg}} = 0.966$ on 689 brain validation pairs (Table S34), and the predicted request reaches 0.366 over 32 regions and 0.448 on the core, against 0.336 and 0.400 for the interval-plus-rate reference and 0.382 and 0.465 for the all-region tabular reference (Table S5). On the 503 validation pairs with a prior visit, graph forecasts exceed the interval-plus-rate reference by +0.030 [+0.006, +0.059] in $R^{(32)}_{\text{chg}}$ (Appendix A.3), and across three matched training seeds cross-region attention raises both request and rendered-change correlations (Table 5(b), Appendix A.2).

Shared region-pair attention bias. A consistent block of the learned region-pair bias recurs across seeds: 285 of 992 directed off-diagonal entries retain their sign across all six training runs under the same configuration, and the median pairwise correlation is 0.503. Figure 3 shows the unanimous pairs: cortical, thalamic and basal-ganglia nodes carry positive biases towards ventricular nodes, and inter-ventricular biases form the strongest negative block (Appendix G.2).

5.3 Request adherence and anatomical rendering

Adherence to predicted anatomy. The renderer follows its request: on test, $\text{MAE}^{(32)}_{\text{adh}}$ is 1.24 against 1.69 pp for copy-forward, a paired reduction of 0.46 [0.39, 0.52] (Appendix B.2). With

Table 4: Brain rendering ablations. Forecaster checkpoints are fixed; retrained variants share the deployed schedule and seed. Test: 669 pairs from 185 subjects; core progression uses 590 pairs. Correlation ± is SD across 32 regions or three core structures; progression ± is pair SD. $^{***}p < 0.001$ for one-sided paired subject-bootstrap tests of full minus variant.

Variant	$R^{(32)}_{\text{chg}} \uparrow$	$R^{(\text{core})}_{\text{chg}} \uparrow$	$\Delta^{(32)}_{\text{RMAE}} \downarrow$	$\Delta^{(\text{core})}_{\text{RMAE}} \downarrow$	$\Delta^{(32)}_{\text{Region}} \downarrow$	$\Delta^{(\text{core})}_{\text{Region}} \downarrow$
GraphMorph (full)	**0.314**±0.094	**0.422**±0.136	**0.800**±0.221	**0.677**±0.447	**1.043**±0.347	0.855±0.590
zero control	0.061***±0.063	0.041***±0.028	1.004±0.335	0.781±0.499	1.189±0.292	0.945±0.590
evolver replaced by identity	0.061***±0.074	-0.015***±0.029	0.976±0.078	0.980±0.101	1.337±0.228	1.357±0.554
latent control signal	0.065***±0.068	0.107***±0.024	1.040±0.142	1.053±0.256	1.401±0.235	1.510±0.590
standard diffusion objective	0.178***±0.127	0.322***±0.179	0.896±0.337	0.685±0.522	1.139±0.328	**0.855**±0.581

Table 5: Spatial forecasting controls. Brain validation, 689 pairs from 186 subjects, training canvas. (a) A shared seed-42 renderer with the deployed graph forecaster or reference requests. ASSD averages (pair, region) cells against measured follow-up labels. (b) Matched forecaster seeds 44–46, mean ± sample SD. Brackets: 95% subject-bootstrap CIs.

	ASSD$^{(32)}$ (mm)↓		Rendered	Raw change
Request source	Request	Rendered	$R^{(32)}_{\text{chg}} \uparrow$	MAE$^{(32)}$ (pp)↓
(a) Content and route through the deployed renderer				
All-region tabular reference → masks	0.486	0.363 [0.349, 0.378]	0.286 [0.260, 0.310]	2.786 [2.677, 2.900]
Interval-plus-rate reference → masks	0.483	0.362 [0.347, 0.377]	0.260 [0.235, 0.288]	2.802 [2.692, 2.919]
Population-trend request → masks	0.480	0.358 [0.343, 0.373]	0.193 [0.163, 0.226]	2.819 [2.701, 2.941]
Graph forecast → masks	0.481	0.359 [0.344, 0.374]	0.280 [0.254, 0.308]	2.746 [2.633, 2.868]
Graph forecast, native request (ours)	**0.474**	**0.349** [0.335, 0.364]	**0.319** [0.293, 0.347]	**2.616** [2.509, 2.730]
Native request − all-region tabular masks	−0.012 [−0.014, −0.010]	−0.014 [−0.015, −0.013]	+0.033 [+0.019, +0.050]	−0.170 [−0.199, −0.142]
Native request − graph masks	−0.007 [−0.009, −0.005]	−0.010 [−0.010, −0.009]	+0.039 [+0.026, +0.053]	−0.131 [−0.152, −0.109]
Graph masks − all-region tabular masks	−0.005 [−0.005, −0.004]	−0.004 [−0.005, −0.003]	−0.006 [−0.024, +0.016]	−0.039 [−0.066, −0.012]
(b) Attention within the native request; mean ± seed SD				
Spatial, self-only	**0.474**±0.000	**0.349**±0.000	0.293±0.002	2.64±0.00
Spatial, full graph	0.474±0.000	0.350±0.000	**0.312**±0.002	**2.62**±0.00
Full graph − self-only	+0.000 [−0.000, +0.001]	+0.000 [+0.000, +0.001]	+0.019 [+0.010, +0.028]	−0.02 [−0.03, −0.01]

a common renderer, the native graph request lowers boundary $\text{ASSD}^{(32)}$ against tabular-derived masks by 0.012 mm at request level and 0.014 mm after rendering, and raises rendered change correlation from 0.286 to 0.319 ($+0.033$ [$+0.019$, $+0.050$]). Through the renderer trained on graph requests, the native spatial request thus carries more rendered change than any mask route on the brain (Table 5; Figure 4).

Changing the request. The request sets the rendered change: holding the renderer fixed, replacing the evolver with the identity reduces $R^{(\text{core})}_{\text{chg}}$ from 0.422 to -0.015 and zero control gives 0.041 (Table 4), while oracle control (F) raises it to 0.704. Through the same renderer, the individualised graph request exceeds the diagnosis-specific population trend in validation $R^{(32)}_{\text{chg}}$ by $+0.126$ [$+0.105$, $+0.151$].

On knee joint space, oracle control (F) raises $R^{(\text{JS})}_{\text{chg}}$ from 0.230 to 0.755 (Table 3). Relative to follow-up-target training, request-target training lowers adherence MAE on both tasks (Table 6) and so buys adherence: on the brain it also raises change correlation by $+0.032$ [$+0.023$, $+0.041$] at no detected cost in forecast MAE, and on knee joint space it trades a forecast-MAE increase of $+0.14$ [$+0.05$, $+0.23$] pp and -0.02 to -0.05 in change correlation for a response slope of 0.442 against 0.203 (Appendix E.3).

Table 6: Renderer supervision on validation. Brain: 689 pairs / 186 subjects, training canvas, 32 regions, seeds 43–45. Knee: 6,176 pairs / 679 patients, joint space, seeds 42–44. Means ± sample SD across seeds. Paired differences: request minus follow-up, 95% cluster-bootstrap CIs (brain subjects, knee patients). Errors use $s = 1$, and response statistics use matched scale sweeps.

Training target	Forecast MAE (pp)↓	Adherence MAE$_{\text{adh}}$ (pp)↓	Response slope	Monotonicity (fraction)
Brain: request-target	**2.613**±0.009	**1.200**±0.016	0.037±0.001	0.800±0.008
follow-up-target	2.623±0.006	1.266±0.005	0.036±0.002	0.750±0.011
paired difference [CI]	−0.010 [−0.021, +0.002]	−0.066 [−0.077, −0.056]	+0.0011 [+0.0007, +0.0174]	+0.0495 [+0.0450, +0.0540]
Knee: request-target	9.20±0.08	**5.16**±0.25	0.442±0.022	0.085±0.025
follow-up-target	**9.06**±0.05	5.76±0.19	0.203±0.037	0.099±0.030
paired difference [CI]	+0.14 [+0.05, +0.23]	−0.60 [−0.70, −0.50]	+0.238 [+0.235, +0.242]	−0.015 [−0.021, −0.008]

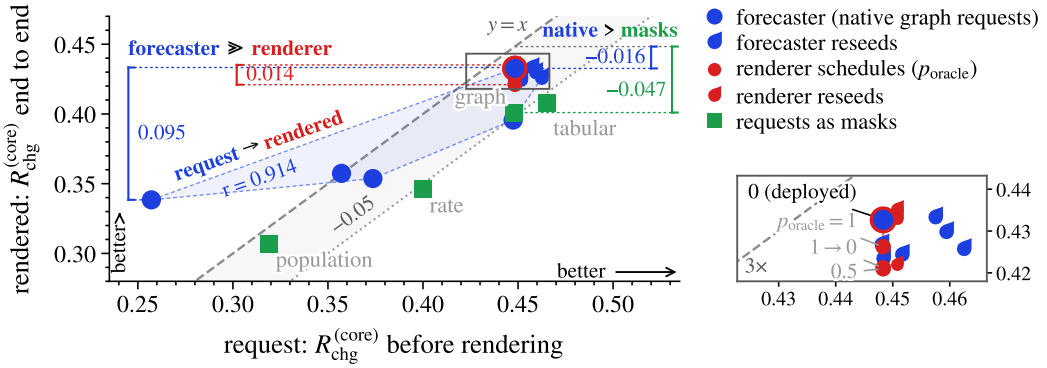

Figure 4: Anatomical request quality before and after rendering. Forecaster variants share one renderer, and renderer variants share one forecaster. Green points show mask-route requests from Table 5(a), whose native request is the deployed arm. Validation: 689 brain pairs, training canvas. Across eleven forecasters (five designs, six reseeds), request and rendered core-change scores correlate at $r = 0.914$ ($\rho = 0.745$). Renderer reseeds are offset horizontally for visibility (Appendix B.4).

Control representation and training objective. After the request itself, the form of the control matters most. With a latent control signal the core correlation falls from 0.422 to 0.107, the standard diffusion objective keeps 0.322 with similar core progression errors, and every ablation raises 32-region progression error (Table 4).

6 Discussion and Conclusion

An explicit, replaceable anatomical forecast separates forecast accuracy from request adherence and shows where longitudinal image forecasting loses its signal. None of the compared pixel-space forecasters recovers core-structure change to the level of elapsed time, whereas a forecast made in anatomical coordinates does. Before rendering, the validation forecast differs from the all-region tabular reference by -0.015 [-0.034, $+0.010$] (Tables S5 and S6). Through one renderer the native spatial request leads every route that converts regional volumes to masks, by $+0.039$ [$+0.026$, $+0.053$] over the same forecast delivered as masks (Table 5). Cross-region attention adds a smaller, consistent gain. Supervising the renderer on its request improves adherence, at a small cost on knee joint space and no detected cost in brain forecast error (Table 6).

Scope. The method applies to identifiable regions with measurable evolution and renders change by deforming baseline intensities.

Reading the references. The tabular reference predicts one scalar per region and returns no image (Section 5.1). The rendered forecast places predicted change at boundaries in the patient's own scan (Figure 2). Routed through the same renderer as masks, the tabular forecast reads 0.286 against 0.319 for the native request (Table 5), so for a future scan the graph route is the better one on the brain. The correlation lead survives an independent segmenter (Table 2). On the knee, attenuated forecasts trade amplitude for ranking, and correlation and amplitude-matched error favour GraphMorph among image forecasters (Table 3).

Limitations. The reproducible attention biases characterise a shared prior and do not isolate individual edges' contributions. Readings are made with a segmenter, and between-segmenter agreement bounds the independent check (Table 2). Most comparisons use single checkpoints, with six forecaster and three renderer reseeds (Appendices A.2 and B.4). On the knee, attenuated forecasts leave joint-space MAE above copy-forward's (9.53 against 8.91 pp), and the tabular reference and its masks lead (Appendix E). Evaluation is retrospective, and prospective accuracy and long-horizon rollout are the next tests.

AI USE STATEMENT

We used ChatGPT and Claude for grammar correction, coding assistance, and exploratory analysis. The authors reviewed all AI-assisted work and take responsibility for the final content.

ETHICS STATEMENT

We used retrospective, de-identified images from ADNI, OASIS-3, AIBL and the Osteoarthritis Initiative (Nevitt et al., 2006). Each source cohort obtained ethical approval and participant consent. No new participant data were collected. The model is intended for research use.

REPRODUCIBILITY STATEMENT

Evaluation protocols, model details, datasets and preprocessing, and training settings are provided in Appendices H–J.

REFERENCES

M. P. Anderson, D. J. Srolovitz, G. S. Grest, and P. S. Sahni. Computer simulation of grain growth—I. Kinetics. *Acta Metallurgica*, 32(5):783–791, 1984. doi: 10.1016/0001-6160(84)90151-2.

Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008. doi: 10.1016/j.media.2007.06.004.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023. doi: 10.1016/j.media.2023.102789.

David Butler, Adrian Hilton, and Gustavo Carneiro. Risk estimation of knee osteoarthritis progression via predictive multi-task modelling from efficient diffusion model using x-ray images. arXiv:2506.14560, 2025.

Víctor M. Campello, Tian Xia, Xiao Liu, Pedro Sanchez, Carlos Martín-Isla, Steffen E. Petersen, Santi Seguí, Sotirios A. Tsaftaris, and Karim Lekadir. Cardiac aging synthesis from cross-sectional data with conditional generative adversarial networks. *Frontiers in Cardiovascular Medicine*, 9:983091, 2022. doi: 10.3389/fcvm.2022.983091.

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. arXiv:1806.07366.

Dexuan Ding, Yuankai Qi, Luping Zhou, Jian Yang, Quan Z. Sheng, and Ming-Hsuan Yang. ProgFormer: Hierarchical voxel diffusion transformer for longitudinal brain MRI prediction. arXiv:2607.27537, 2026.

Nico Albert Disch, Yannick Kirchhoff, Robin Peretzke, Maximilian Rokuss, Saikat Roy, Constantin Ulrich, David Zimmerer, and Klaus Maier-Hein. Temporal flow matching for learning spatio-temporal trajectories in 4D longitudinal medical imaging. arXiv:2508.21580, 2025.

Mengjin Dong, Long Xie, Sandhitsu R. Das, Jiancong Wang, Laura E. M. Wisse, Robin deFlores, David A. Wolk, and Paul A. Yushkevich. Deepatrophy: Teaching a neural network to detect progressive changes in longitudinal MRI of the hippocampal region in Alzheimer's disease. *NeuroImage*, 243:118514, 2021. doi: 10.1016/j.neuroimage.2021.118514.

Kathryn A. Ellis, Ashley I. Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T. Lautenschlager, Nat Lenzo, Ralph N. Martins, Paul Maruff, Colin Masters, Andrew Milner, Kerryn Pike, Christopher Rowe, Greg Savage, Cassandra Szoeke, Kevin Taddei, Victor Villemagne, Michael Woodward, and David Ames. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *International Psychogeriatrics*, 21(4): 672–687, 2009. doi: 10.1017/S1041610209009405.

Ali Farki, Elaheh Moradi, Deepika Koundal, and Jussi Tohka. Forecasting future anatomies: Longitudinal brain MRI-to-MRI prediction. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2026. doi: 10.1109/ISBI61048.2026.11515462.

Yuxiang Feng, Juncheng Wang, Chao Xu, Wenlong Hou, Huihan Wang, Yijie Qian, Yang Liu, Baigui Sun, Yong Liu, and Shujun Wang. Progression as latent drift: Generative forecasting of slow-evolving pathologies. arXiv:2607.08270, 2026.

Hubert M. Fonteijn, Marc Modat, Matthew J. Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z. Hobbs, Rachael I. Scahill, Sarah J. Tabrizi, Sébastien Ourselin, Nick C. Fox, and Daniel C. Alexander. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012. doi: 10.1016/j.neuroimage.2012.01.062.

Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack Jr, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010. doi: 10.1038/nrneurol.2009.215.

Jingru Fu, Yuqi Zheng, Neel Dey, Daniel Ferreira, and Rodrigo Moreno. Synthesizing individualized aging brains in health and disease with generative models and parallel transport. *Medical Image Analysis*, 105:103669, 2025. doi: 10.1016/j.media.2025.103669.

Yifeng Gao, Zhen Zhou, Bing Zhang, Saidi Guo, Kairui Bo, Shuang Li, Nan Zhang, Hui Wang, Guang Yang, Heye Zhang, Tong Liu, and Lei Xu. Deep learning-based prognostic model using non-enhanced cardiac cine MRI for outcome prediction in patients with heart failure. *European Radiology*, 33(11):8203–8213, 2023. doi: 10.1007/s00330-023-09785-9.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. arXiv:1406.2661.

Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. MAISI: Medical AI for synthetic imaging. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, 2025. doi: 10.1109/WACV61041.2025.00435.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219: 117012, 2020. doi: 10.1016/j.neuroimage.2020.117012.

Jiaping Hu, Chuanyang Zheng, Qingling Yu, Lijie Zhong, Keyan Yu, Yanjun Chen, Zhao Wang, Bin Zhang, Qi Dou, and Xiaodong Zhang. Deepkoa: a deep-learning model for predicting progression in knee osteoarthritis using multimodal magnetic resonance images from the Osteoarthritis Initiative. *Quantitative Imaging in Medicine and Surgery*, 13(8):4852–4866, 2023. doi: 10.21037/qims-22-1251.

David J. Hunter and Sita Bierma-Zeinstra. Osteoarthritis. *The Lancet*, 393(10182):1745–1759, 2019. doi: 10.1016/S0140-6736(19)30417-9.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017. doi: 10.1109/CVPR.2017.632.

Bailiang Jian, Jiazhen Pan, Yitong Li, Fabian Bongratz, Ruochen Li, Daniel Rueckert, Benedikt Wiestler, and Christian Wachinger. TimeFlow: Temporal conditioning for longitudinal brain MRI registration and aging analysis. arXiv:2501.08667, 2025.

Jaivardhan Kapoor, Jakob H. Macke, and Christian F. Baumgartner. MRExtrap: Longitudinal aging of brain MRIs using linear modeling in latent space. arXiv:2508.19482, 2025.

Bishesh Khanal, Nicholas Ayache, and Xavier Pennec. Simulating longitudinal brain MRIs with known volume changes and realistic variations in image intensity. *Frontiers in Neuroscience*, 11: 132, 2017. doi: 10.3389/fnins.2017.00132.

Boah Kim, Inhwa Han, and Jong Chul Ye. DiffuseMorph: Unsupervised deformable image registration using diffusion model. In *European Conference on Computer Vision (ECCV)*, pp. 347–364, 2022. doi: 10.1007/978-3-031-19821-2_20.

Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 88–98, 2024. doi: 10.1007/978-3-031-72104-5_9.

Pamela J. LaMontagne, Tammie L. S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*, 2019. doi: 10.1101/2019.12.13.19014902.

Binxu Li, Wei Peng, Mingjie Li, Ehsan Adeli, and Kilian M. Pohl. Integrating anatomical priors into a causal diffusion model. arXiv:2509.09054, 2025. URL https://arxiv.org/abs/2509.09054.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2210.02747.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. TADM: Temporally-aware diffusion model for neurodegenerative progression on brain MRI. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15002 of *Lecture Notes in Computer Science*, pp. 444–453. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72069-7_42.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. Temporally-aware diffusion model for brain progression modelling with bidirectional temporal regularisation. arXiv:2509.03141, 2025.

Chen Liu, Ke Xu, Liangbo L. Shen, Guillaume Huguet, Zilong Wang, Alexander Tong, Danilo Bzdok, Jay Stewart, Jay C. Wang, Lucian V. Del Priore, and Smita Krishnaswamy. ImageFlowNet: Forecasting multiscale image-level trajectories of disease progression with irregularly-sampled longitudinal medical images. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025. Oral. arXiv:2406.14794.

Wenbing Lv, Junyi Peng, Jiaping Hu, Yijun Lu, Zidong Zhou, Hui Xu, Kongzai Xing, Xiaodong Zhang, and Lijun Lu. LMSST-GCN: Longitudinal MRI sub-structural texture guided graph convolution network for improved progression prediction of knee osteoarthritis. *Computer Methods and Programs in Biomedicine*, 261:108600, 2025. doi: 10.1016/j.cmpb.2025.108600.

Răzvan V. Marinescu, Neil P. Oxtoby, Alexandra L. Young, Esther E. Bron, Arthur W. Toga, Michael W. Weiner, Frederik Barkhof, Nick C. Fox, Polina Golland, Stefan Klein, and Daniel C. Alexander. TADPOLE challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data. In *Predictive Intelligence in Medicine*, volume 11843 of *Lecture Notes in Computer Science*, pp. 1–10, 2019. doi: 10.1007/978-3-030-32281-6_1.

Gerard Martí-Juan, Marco Lorenzi, and Gemma Piella. MC-RVAE: Multi-channel recurrent variational autoencoder for multimodal Alzheimer's disease progression modelling. *NeuroImage*, 268: 119892, 2023. doi: 10.1016/j.neuroimage.2023.119892.

Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M. Jorge Cardoso, Marc Modat, Sébastien Ourselin, and Lauge Sørensen. Training recurrent neural networks robust to incomplete data: Application to Alzheimer's disease progression modeling. *Medical Image Analysis*, 53:39–46, 2019. doi: 10.1016/j.media.2019.01.004.

Andrea Moschetto, Lemuel Puglisi, Alec Sargood, Pierluigi Dell'Acqua, Francesco Guarnera, Sebastiano Battiato, and Daniele Ravì. Benchmarking GANs, diffusion models, and flow matching for T1w-to-T2w MRI translation. arXiv, 2025. arXiv:2507.14575.

Susanne G. Mueller, Michael W. Weiner, Leon J. Thal, Ronald C. Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, and Laurel Beckett. The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869–877, 2005. doi: 10.1016/j.nic.2005.09.008.

W. W. Mullins. Two-dimensional motion of idealized grain boundaries. *Journal of Applied Physics*, 27(8):900–904, 1956.

Michael C. Nevitt, David T. Felson, and Gayle Lester. The Osteoarthritis Initiative: Protocol for the cohort study. Study protocol, Osteoarthritis Initiative, 2006. URL `https://nda.nih.gov/static/docs/StudyDesignProtocolAndAppendices.pdf`.

Minh Nguyen, Tong He, Lijun An, Daniel C. Alexander, Jiashi Feng, and B. T. Thomas Yeo. Predicting Alzheimer's disease progression using deep recurrent neural networks. *NeuroImage*, 222: 117203, 2020. doi: 10.1016/j.neuroimage.2020.117203.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Charles G. Peterfy, Erika Schneider, and Michael Nevitt. The osteoarthritis initiative: report on the design rationale for the magnetic resonance imaging protocol for the knee. *Osteoarthritis and Cartilage*, 16(12):1433–1441, 2008. doi: 10.1016/j.joca.2008.06.016.

Steve Plimpton, Corbett Battaile, Mike Chandross, Liz Holm, Aidan Thompson, Veena Tikare, Greg Wagner, Ed Webb, Xiaowang Zhou, Cristina Garcia Cardona, and Alex Slepoy. Crossing the mesoscale no-man's land via parallel kinetic Monte Carlo. Technical Report SAND2009-6226, Sandia National Laboratories, 2009.

Guilherme Pombo, Robert Gray, M. Jorge Cardoso, Sebastien Ourselin, Geraint Rees, John Ashburner, and Parashkev Nachev. Equitable modelling of brain imaging by counterfactual augmentation with morphologically constrained 3D deep generative models. *Medical Image Analysis*, 84:102723, 2023. doi: 10.1016/j.media.2022.102723.

Fabi Prezja, Leevi Annala, Sampsa Kiiskinen, Suvi Lahtinen, Timo Ojala, and Paavo Nieminen. Generating synthetic past and future states of knee osteoarthritis radiographs using cycle-consistent generative adversarial neural networks. *Computers in Biology and Medicine*, 187: 109785, 2025. doi: 10.1016/j.compbiomed.2025.109785.

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. doi: 10.1016/j.media.2025.103734.

Daniele Ravì, Stefano B. Blumberg, Silvia Ingala, Frederik Barkhof, Daniel C. Alexander, and Neil P. Oxtoby. Degenerative adversarial neuroimage nets for brain scan simulations: Application in ageing and dementia. *Medical Image Analysis*, 75:102257, 2022. doi: 10.1016/j.media.2021.102257.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, 2022. doi: 10.1109/CVPR52688.2022.01042.

Ahmed M. Salih, Esmeralda Ruiz Pujadas, Víctor M. Campello, Celeste McCracken, Nicholas C. Harvey, Stefan Neubauer, Karim Lekadir, Thomas E. Nichols, Steffen E. Petersen, and Zahra Raisi-Estabragh. Image-based biological heart age estimation reveals differential aging patterns across cardiac chambers. *Journal of Magnetic Resonance Imaging*, 58(6):1797–1812, 2023. doi: 10.1002/jmri.28675.

Benoît Sauty and Stanley Durrleman. Progression models for imaging data with longitudinal variational auto encoders. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13431 of *Lecture Notes in Computer Science*, pp. 3–13. Springer, Cham, 2022. doi: 10.1007/978-3-031-16431-6_1.

Pengwei Sun, Wei Peng, Lun Yu Li, Yixin Wang, and Kilian M. Pohl. Evaluation of 3D counterfactual brain MRI generation. arXiv:2508.02880, 2025.

Jean-Philippe Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998. doi: 10.1016/S1361-8415(98)80022-4.

Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11:9068, 2021. doi: 10.1038/s41598-021-87564-6.

Tom Vercauteren, Xavier Pennec, Aymeric Perchant, and Nicholas Ayache. Diffeomorphic demons: efficient non-parametric image registration. *NeuroImage*, 45(1 Suppl):S61–S72, 2009. doi: 10.1016/j.neuroimage.2008.10.040.

Cheng Wan, Bahram Jafrasteh, Ehsan Adeli, Miaomiao Zhang, and Qingyu Zhao. Anatomically guided latent diffusion for brain MRI progression modeling. arXiv:2601.14584, 2026.

Wenhua Wu, Kun Hu, Wenxi Yue, Wei Li, Milena Simic, Changyang Li, Wei Xiang, and Zhiyong Wang. Identity-consistent diffusion network for grading knee osteoarthritis progression in radiographic imaging. arXiv:2407.21381, 2024.

Tian Xia, Agisilaos Chartsias, Chengjia Wang, and Sotirios A. Tsaftaris. Learning to synthesise the ageing brain without longitudinal data. *Medical Image Analysis*, 73:102169, 2021. doi: 10.1016/j.media.2021.102169.

Hao Yang, Tao Tan, Shuai Tan, Weiqin Yang, Kunyan Cai, Calvin Chen, and Yue Sun. Mambacontrol: Anatomy graph-enhanced mamba controlnet with fourier refinement for diffusion-based disease trajectory prediction. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 4423–4428. IEEE, 2025. doi: 10.1109/SMC58881.2025.11343720.

Alexandra L. Young, Răzvan V. Marinescu, Neil P. Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C. Firth, David M. Cash, David L. Thomas, Katrina M. Dick, Jorge Cardoso, John van Swieten, Barbara Borroni, Daniela Galimberti, Mario Masellis, Maria Carmela Tartaglia, James B. Rowe, Caroline Graff, Fabrizio Tagliavini, Giovanni B. Frisoni, Robert Laforce, Elizabeth Finger, Alexandre de Mendonça, Sandro Sorbi, Jason D. Warren, Sebastian Crutch, Nick C. Fox, Sébastien Ourselin, Jonathan M. Schott, Jonathan D. Rohrer, and Daniel C. Alexander. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature Communications*, 9(1):4273, 2018. doi: 10.1038/s41467-018-05892-0.

Alexandra L. Young, Neil P. Oxtoby, Sara Garbarino, Nick C. Fox, Frederik Barkhof, Jonathan M. Schott, and Daniel C. Alexander. Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nature Reviews Neuroscience*, 25(2):111–130, 2024. doi: 10.1038/s41583-023-00779-6.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3813–3824, 2023. doi: 10.1109/ICCV51070.2023.00355.

SUPPLEMENTARY RESULTS AND TECHNICAL DETAILS

A. Forecast contributions and input controls	15	G. Representation and graph diagnostics	37
B. Request adherence and rendering	19	H. Evaluation protocol	39
C. Measurement reliability and robustness	22	I. Method and reproducibility details	45
D. Brain MRI benchmark	28	I.7. Training settings and reproducibility	48
E. Knee radiographs	32	J. Data and cohort construction	49
F. Grain-growth extension	35	K. Extended related work	51

The six questions the paper answers, and where the evidence for each is:

Question	Main text	Appendix tables	Figures
Q1 How accurately does the pipeline predict anatomical change relative to image forecasters?	Section 5.1; Tables 1, 3	Tables S23, S25, S26	Figures 2, S6, S5
Q2 Is the render faithful to its request, and does the prediction gain reach the image?	Section 5.3; Tables 4, 6	Tables S9, S11, S13, S14	Figures 4, S1
Q3 How do history, cross-region interaction and request representation affect forecasts and rendered anatomy?	Sections 5.2, 5.3 and Table 5	Tables S2, S3, S4, S5, S7, S1	Figures 3, S7
Q4 Is the result robust to the segmenter, to calibration and magnitude, across strata, at boundaries and under amplitude matching?	Section 5.1, Table 2	Tables S20, S22, S15, S17, S18, S19	Figures S4, S3
Q5 Does the pipeline transfer to knee radiographs and grain growth?	Section 5.1; Table 3; Section 6	Tables S27, S30, S31, S32	—
Q6 What protocol, budgets and settings reproduce it?	Section 4	Tables S35, S36, S37	Figure 1

Figures and tables use the prefix S.

A FORECAST CONTRIBUTIONS AND INPUT CONTROLS

These analyses separate the effects of input history, graph interaction and request representation. Request-level scores evaluate decoded anatomy; end-to-end scores evaluate anatomy measured from the rendered image. Each comparison identifies its split, canvas and source of uncertainty; baseline definitions and metric formulas are in Appendix H.2.

A.1 BRAIN METHODS AND INPUT CONTROLS

The compared methods differ in what they see of the subject: GraphMorph, BrLP and Δ-LFM receive history, the MambaControl backbone, CardiacAging and ImageFlowNet the baseline and the elapsed interval, and DiffuseMorph and the four image-to-image baselines the baseline image alone. GraphMorph receives baseline anatomy, the input-visit diagnosis (Appendix I.7), available anatomical history, age and the elapsed interval. BrLP is conditioned on the target age, on follow-up covariates estimated by its own disease-course stage from the subject's history, and on the target visit's diagnosis label. Δ-LFM receives the baseline latent and up to two prior latents with their intervals, plus the diagnosis change between the two visits, measured at the target visit and non-zero on 36 of the 669 test pairs. The MambaControl backbone receives the baseline latent, the elapsed

interval and the target age; the ControlNet slots trained on the follow-up diagnosis and five follow-up volumes carry baseline values at inference and the sex slot a constant, which differs from its training conditioning. Its test-time free parameters are frozen on validation (Appendix I.6).

Image comparators were selected by change correlation on one prespecified validation pair per subject (brain) or patient (knee): Δ-LFM and Pix2Pix, respectively. Comparisons use method-specific inputs and training budgets; Table S1 reports input and objective controls.

Table S1: Brain input and objective controls on 669 test pairs from 185 subjects. Metrics average 32 regions; MAE is in percentage points. Brackets give 95% paired subject-bootstrap CIs. Differences are first minus second.

Configuration	$R^{(32)}_{\text{chg}}$	$\text{MAE}^{(32)}$
Full graph, seed 44	0.313	2.67
No correlation loss	0.310	2.68
No history inputs	0.224	2.75
Δ-LFM	0.230	5.90
Paired differences: first minus second		
No correlation loss − full graph	-0.003 $[-0.010, +0.005]$	$+0.016$ $[+0.006, +0.026]$
No history inputs − Δ-LFM	-0.006 $[-0.036, +0.035]$	-3.16 $[-3.29, -3.03]$

A.2 Forecast accuracy and what carries it

The deployed forecaster predicts decoded anatomical change before rendering at $R^{(32)}_{\text{chg}} = 0.366$ and $R^{(\text{core})}_{\text{chg}} = 0.448$ (Table S2); the seed-42 checkpoint was selected by core correlation before reseeding.

The deployed request improves core correlation and MAE over the Leaspy disease-course reference without images (0.448 versus 0.433; 2.43 versus 2.71 pp) and 32-region MAE over predicting no change (2.47 versus 2.81 pp), against a measurement ceiling of 0.729 ± 0.060 on the core and a Stage 1 round trip of the true future graph at 0.969 over 32 regions and 0.984 on the core. Among the three core structures, its correlation exceeds the merged-region tabular reference only for the lateral ventricle. All six reseeds exceed the Baseline row in core correlation (Appendix I.7).

Table S2: Brain forecaster design on 689 validation pairs from 186 subjects. Correlation and MAE are means $\pm$ SD across 32 regions or three core structures, except the six-reseed row, which gives SD across seeds 44–49. MAE is in percentage points.

Arm	$R^{(32)}_{\text{chg}} \uparrow$	$R^{(\text{core})}_{\text{chg}} \uparrow$	$\text{MAE}^{(32)} \downarrow$	$\text{MAE}^{(\text{core})} \downarrow$
All-region tabular reference	0.382 ± 0.094	0.465 ± 0.093	—	—
Baseline: 32-region graph, stochastic 8-region decode	0.213 ± 0.136	0.357 ± 0.162	2.59 ± 1.24	2.50 ± 0.59
Decode supervision: all 32 regions at every step	0.217 ± 0.129	0.257 ± 0.192	2.61 ± 1.27	2.56 ± 0.61
Decode supervision: the core-6 regions only	—	0.374 ± 0.157	—	2.49 ± 0.60
Deployed: all regional previous-interval rates	0.366 ± 0.090	$\mathbf{0.448} \pm 0.110$	2.47 ± 1.20	$\mathbf{2.43} \pm 0.58$
Deployed − per-step Δt scaling	$\mathbf{0.377} \pm 0.084$	0.447 ± 0.100	$\mathbf{2.46} \pm 1.20$	2.44 ± 0.59
Deployed, six reseeds (44–49)	0.368 ± 0.005	0.455 ± 0.006	2.47 ± 0.00	2.43 ± 0.01

On validation, hippocampus/amygdala/lateral-ventricle correlations are $0.726/0.657/0.804$ for the measurement ceiling, $0.972/0.981/0.999$ for the Stage 1 round trip, $0.393/0.382/0.588$ for the merged-region tabular reference and $0.387/0.355/0.603$ for the deployed request. Six forecaster reseeds span 0.362–0.377 over 32 regions and 0.448–0.463 on the core.

Historical inputs and design comparisons. Previous-interval rates carry the tabular reference as well: adding them to a boosted predictor with baseline volumes and demographics raises 32-region validation correlation from 0.264 to 0.373, and per-region selection between ridge and boosting improves this reference further (Table S2).

Removing per-step Δt scaling leaves core correlation nearly unchanged and raises the 32-region score; full-32 decode supervision gives lower core correlation (Table S2). These design arms also differ in optimisation (Appendix I.7).

Graph interaction and request controls. Cross-region attention in the evolver raises both request and rendered-change correlations on all three matched seeds. With the renderer frozen, end-to-

end core correlation is 0.429 ± 0.004 with cross-region attention and 0.401 ± 0.003 with self-only attention (Table S3; paired 32-region render contrast, self-only minus full graph, -0.019 [-0.028, -0.010]).

Table S3: Cross-region attention on 689 brain validation pairs from 186 subjects, training canvas. Regional correlations are averaged over 32 regions or three core structures. Values are means ± SD across matched seeds 44–46.

	Request		Render	
Attention	$R^{(32)}_{\text{chg}}$	$R^{(\text{core})}_{\text{chg}}$	$R^{(32)}_{\text{chg}}$	$R^{(\text{core})}_{\text{chg}}$
Full	**0.366** ± 0.002	**0.456** ± 0.007	**0.312** ± 0.002	**0.429** ± 0.004
Self-only	0.350 ± 0.002	0.431 ± 0.004	0.293 ± 0.002	0.401 ± 0.003

The native spatial route improves rendered-change correlation. The renderer, trained on graph requests, follows native requests more closely than the same predicted percentage changes re-expressed as masks on measured baseline anatomy (Table S4), and through the same frozen renderer every mask route scores lower (Tables 5 and S5). The decoded graph forecast and its decoded identity anchor have the same request-boundary ASSD (0.474 mm). The native request's margin over the diagnosis-specific population-trend request, $+0.126$ [$+0.105$, $+0.151$] over all pairs, is $+0.159$ [$+0.134$, $+0.188$] in pairs with a prior interval and $+0.013$ [-0.009, $+0.032$] on first pairs, with lower $\text{MAE}^{(32)}$ (-0.203 [-0.235, -0.172] pp over all pairs). With the graph forecast delivered through the same mask route as the population request (post hoc), the all-pairs margin is $+0.087$ [$+0.068$, $+0.108$]. Population and direct-history requests are defined in Appendix H.2.

Table S4: Request representation on 689 brain validation pairs from 186 subjects, training canvas, seed 42. Differences are mask route minus native graph, with paired subject-bootstrap 95% CIs. Core6 retains six separate regions.

Metric	Mask route	Native graph	Difference [95% CI]
$\text{MAE}^{(32)}_{\text{adh}}$ (pp)	1.411	1.193	$+0.217$ [$+0.198$, $+0.238$]
$\text{MAE}^{(\text{core6})}_{\text{adh}}$ (pp)	1.240	0.874	$+0.366$ [$+0.323$, $+0.412$]
$\text{ASSD}^{(32)}$ (mm)	0.359	0.349	$+0.010$ [$+0.009$, $+0.010$]
$\text{ASSD}^{(\text{core6})}$ (mm)	0.346	0.332	$+0.014$ [$+0.013$, $+0.016$]
$\text{SDice}^{(32)}_{1.5}$	0.989	0.990	-0.001 [-0.001, -0.000]

Table S5: Forecast content and request route on 689 brain validation pairs from 186 subjects, training canvas. Native and mask routes share the seed-42 renderer. Request seed means average metrics across seeds 44–46. Correlation and MAE average regions; brackets give 95% subject-bootstrap CIs.

Content and route	$R^{(32)}_{\text{chg}}$ ↑	$R^{(\text{core})}_{\text{chg}}$ ↑	$\text{MAE}^{(32)}$ (pp)↓
Request level, before rendering			
Population trend (diagnosis-specific)	0.179 [0.145, 0.216]	0.319 [0.253, 0.389]	2.624 [2.506, 2.747]
Interval + rate (history only)	0.336 [0.310, 0.365]	0.400 [0.343, 0.456]	2.527 [2.416, 2.648]
All-region tabular	0.382 [0.354, 0.410]	0.465 [0.409, 0.520]	2.518 [2.411, 2.633]
Graph forecast (seeds 44–46)	0.366 [0.342, 0.396]	0.456 [0.392, 0.515]	2.472 [2.363, 2.592]
Graph forecast (seed 42)	0.366 [0.342, 0.396]	0.448 [0.386, 0.506]	2.472 [2.363, 2.590]
Mask: through the fixed renderer			
Population trend	0.193 [0.163, 0.226]	0.307 [0.234, 0.382]	2.819 [2.701, 2.941]
Interval + rate	0.260 [0.235, 0.288]	0.346 [0.291, 0.398]	2.802 [2.692, 2.919]
All-region tabular	0.286 [0.260, 0.310]	0.408 [0.350, 0.466]	2.786 [2.677, 2.900]
Graph forecast (seed 42)	0.280 [0.254, 0.308]	0.401 [0.345, 0.454]	2.746 [2.633, 2.868]
Native: through the fixed renderer			
Graph forecast (seed 42)	**0.319** [0.293, 0.347]	**0.433** [0.370, 0.495]	**2.616** [2.509, 2.730]

A.3 Available history and tabular references

A prior visit carries the forecaster's lead. With one, the deployed pipeline exceeds the elapsed-interval reference in test correlation; without one, the core correlations are close (Table S7; the strata also differ in subjects and follow-up intervals). On the shortest of three interval terciles (223 test pairs) the elapsed-interval reference has the higher core correlation (0.247 versus 0.244). On the full test cohort, the all-region tabular reference is stronger than the deployed pipeline on both

Table S6: Paired component comparisons in change correlation: 32 brain regions or knee joint space. Differences are first minus second; samples and settings are specified by row. Brackets give 95% subject-bootstrap CIs, using patients for knee.

Contrast (first − second)	ΔR_{chg} [95% CI]	Sample and location
Graph forecaster		
Graph forecast (seed mean) − interval-plus-rate reference (request level)	+0.030 [+0.006, +0.059]	503 validation pairs with a prior visit; Appendix A.3
Graph forecast (seed mean) − all-region tabular reference (request level)	−0.015 [−0.034, +0.010]	689 pairs; Table S5
Full graph − self-only attention (rendered)	+0.019 [+0.010, +0.028]	689 pairs, seeds 44–46; Table 5(b)
Spatial request (rendered)		
Native request − graph forecast as masks	+0.039 [+0.026, +0.053]	689 pairs; Table S5
Graph forecast as masks − all-region tabular reference as masks	−0.006 [−0.024, +0.016]	689 pairs; Table S5
Native request − population-trend request as masks	+0.126 [+0.105, +0.151]	689 pairs; Appendix A.2
Knee: tabular masks − graph request ($R^{(\text{JS})}_{\text{chg}}$)	+0.063 [+0.041, +0.084]	6,176 pairs / 679 patients; Appendix E.3
Renderer training target (rendered, matched seeds)		
Request-target − follow-up-target training	+0.032 [+0.023, +0.041]	689 pairs, seeds 43–45; Appendix B.2

aggregates, and the interval-plus-rate reference is stronger over 32 regions and similar on the core (Table S8; reference configurations in Appendix H.2).

Table S7: Brain test results by available history, on 669 pairs from 185 subjects; stratum sizes count pairs. Values are means ± SD across 32 regions or three core structures. MAE is in percentage points.

Method / reference	$R^{(32)}_{\text{chg}}$	$R^{(\text{core})}_{\text{chg}}$	$\text{MAE}^{(\text{core})}$
Has prior visit; 484 pairs			
GraphMorph (ours)	**0.332**±0.083	**0.432**±0.096	**2.63**±0.71
Δ-LFM	0.249±0.138	0.169±0.095	4.68±0.78
CardiacAging	0.087±0.108	0.247±0.135	3.58±1.14
Elapsed interval	0.161±0.127	0.278±0.180	—
Copy-forward	—	—	3.60±1.46
No prior visit; 185 pairs			
GraphMorph (ours)	**0.233**±0.176	**0.356**±0.254	**2.74**±0.76
Δ-LFM	0.174±0.160	0.312±0.242	5.24±1.36
CardiacAging	0.137±0.155	0.297±0.242	3.26±1.00
Elapsed interval	0.209±0.179	0.337±0.238	—
Copy-forward	—	—	3.26±1.35

Table S8: Full-cohort tabular references on 689 brain validation pairs from 186 subjects and 669 test pairs from 185 subjects. Values are means ± SD across 32 regions or three core structures. Feature definitions are in Appendix H.2.

Reference	Valid $R^{(32)}_{\text{chg}}$	Valid $R^{(\text{core})}_{\text{chg}}$	Test $R^{(32)}_{\text{chg}}$	Test $R^{(\text{core})}_{\text{chg}}$
Elapsed interval (validation-selected fit)	0.176±0.140	0.316±0.189	0.174±0.148	0.324±0.177
Interval-plus-rate	0.336±0.088	0.400±0.134	0.334±0.093	0.395±0.134
All-region tabular	0.382±0.094	0.465±0.093	0.369±0.100	0.493±0.085

The validation comparisons hold the evaluated pairs fixed. On the 503 pairs (155 subjects) with a prior interval, graph requests improve $R^{(32)}_{\text{chg}}$ over the interval-plus-rate reference by $+0.030$ [$+0.006$, $+0.059$] ($R^{(\text{core})}_{\text{chg}}$ $+0.071$ [$+0.029$, $+0.110$]) and differ from the all-region tabular reference by -0.007 [-0.028, $+0.021$]; the history inputs are worth $+0.209$ [$+0.166$, $+0.256$] over a no-history model (seed 44), and directly extrapolating each region's previous rate scores $R^{(32)}_{\text{chg}}$ -0.169 [-0.236, -0.130]. On the 186 first pairs, the all-region tabular reference has the higher 32-region correlation (graph minus tabular -0.041 [-0.076, -0.002]) and the history inputs change the forecast by $+0.007$ [-0.012, $+0.023$]. In the AD stratum (137 pairs / 55 subjects), the rendered forecast leads Δ-LFM by $+0.099$ [$+0.022$, $+0.178$] in $R^{(32)}_{\text{chg}}$, whereas at request level the graph seed mean differs from the all-region tabular forecast by -0.025 [-0.055, $+0.014$].

A.4 A SEMANTICS-FREE REQUEST

A request that names no structure forecasts far worse than the named request: end-to-end $R^{(32)}_{\text{chg}}$ is 0.087 against 0.314 on test. The grid replaces the 32 named SynthSeg structures with each subject's

first-visit brain bounding box divided $4 \times 4 \times 4$; the $K{=}56$ cells holding at least 500 brain voxels in 99% of training visits are kept, and a cell's label map is the brain mask within it (construction details in Appendix I.7). All three stages were retrained once at seed 42 with the deployed configuration and renderer backbone, and the renders are read through SynthSeg on the 32 named regions, on the 689 validation pairs (default canvas) and on the 669 test pairs (training canvas).

Over the 56 grid cells, the decoded forecast correlates with true cell change at a mean r of 0.272 on validation, against 0.044 for the elapsed interval and 0.963 for the decoded measured follow-up. On the named structures (95% pair-bootstrap intervals) the grid request reads 0.086 [0.068, 0.102] over the 32 regions and 0.073 [0.029, 0.116] on the core on validation, against 0.280 and 0.417 for the named request through its own renderer, and 0.087 [0.054, 0.114] and 0.028 [-0.027, 0.084] on test against 0.314 and 0.422. The comparison also changes region count and Stage 1 batch size.

A.5 A FORECAST FAILURE

The selected failure is a forecasting error, not a rendering error: the render follows a request for substantially more change than occurred (Figure S6c). The subject's mean core request error is 8.3 pp, against a median of 2.7 across the 127 eligible test subjects; its first and last intervals contribute errors of 17.3 and 16.3 pp.

B REQUEST ADHERENCE AND RENDERING

B.1 THE CONTROL SIGNAL

Decoded occupancy, not the latent increment, is the usable control signal (Eq. 3). On 689 validation pairs with the deployed checkpoints and no rendering, the predicted and true increments have median cosine 0.005 across 32 regions and a maximum of 0.025 in the right caudate, yet decoding the predicted increment gives mean change correlation 0.366 (range 0.178–0.601), and decoding the measured follow-up graph gives median 0.972 with no region below 0.899.

B.2 REQUEST AGREEMENT AND REGISTRATION CONTROLS

Test renders lie closer to their request than copy-forward, the standard-objective renderer and latent control. Table S9 separates request accuracy from the render's agreement with its request: the decoded forecast reaches $R^{(32)}_{\mathrm{chg}} = 0.370$ against its own decoded baseline and 0.299 against the measured baseline, and the rendered forecast reads 0.314. The baseline definition changes how reconstruction error enters the first comparison.

Table S9: Request agreement and forecast accuracy on 669 brain test pairs from 185 subjects. ± denotes SD across regions; core6 retains six separate labels. MAE is in percentage points. The two decoded-request rows use their own decoded baseline or the common measured baseline.

	Reference-request agreement			Measured follow-up			
Arm	$\mathrm{MAE}^{(32)}_{\mathrm{adh}}$	$\mathrm{MAE}^{(\mathrm{core6})}_{\mathrm{adh}}$	$R^{(32)}_{\mathrm{adh}}$	$\mathrm{MAE}^{(32)}$	$\mathrm{MAE}^{(\mathrm{core6})}$	$R^{(32)}_{\mathrm{chg}}$	$R^{(\mathrm{core6})}_{\mathrm{chg}}$
Decoded request, q^{decoded}	—	—	—	2.54±1.32	2.97±0.78	0.370±0.083	0.427±0.121
Decoded request, q^{common}	—	—	—	2.69±1.41	3.09±0.76	0.299±0.104	0.364±0.165
GraphMorph (ours)	**1.24**±0.62	**0.90**±0.30	**0.607**±0.150	**2.67**±1.29	**3.02**±0.79	**0.314**±0.094	**0.399**±0.145
Latent control signal	2.48±1.79	2.82±1.78	0.051±0.084	3.44±1.93	4.02±1.71	0.065±0.068	0.104±0.069
Standard diffusion objective	2.01±1.41	1.71±0.20	0.216±0.238	3.04±1.68	3.21±0.74	0.178±0.127	0.314±0.181
Copy-forward ($\hat{F} = M$)	1.69±1.31	2.45±1.28	—	2.94±1.70	3.86±1.31	—	—

Table S10: Paired MAE reductions on 669 brain test pairs from 185 subjects: comparator minus GraphMorph. Reference agreement and end-to-end errors are separate endpoints. Brackets give 95% paired subject-bootstrap CIs; MAE is in percentage points and core6 retains six separate labels.

Comparator	Metric	$\mathrm{MAE}^{(32)}$	$\mathrm{MAE}^{(\mathrm{core6})}$
Copy-forward	Reference agreement	+0.46 [+0.39, +0.52]	+1.55 [+1.39, +1.72]
Standard diffusion objective	Reference agreement	+0.77 [+0.73, +0.80]	+0.81 [+0.75, +0.87]
Latent control signal	Reference agreement	+1.24 [+1.19, +1.29]	+1.92 [+1.79, +2.07]
Copy-forward	End-to-end	+0.28 [+0.22, +0.34]	+0.84 [+0.69, +0.99]

For the same anatomical forecast, label-driven SyN gives lower change correlations and larger request errors than the learned renderer at nearly the same raw forecast MAE and boundary distance (Table S11); the hard-mask reference gives the same request-error ordering ($\mathrm{MAE}^{(32)}_{\mathrm{adh}}$ $+0.144$ [$+0.127$, $+0.161$] pp against $+0.151$ [$+0.134$, $+0.169$] with the soft reference).

Table S11: Forecast rendering, reconstruction anchor and SyN registration on 689 brain validation pairs from 186 subjects. Differences are first minus second; brackets give paired subject-bootstrap 95% CIs. Core denotes three merged structures, and MAE is in percentage points.

Metric	Renderer	SyN forecast	Anchor	SyN forecast − renderer	SyN forecast − anchor
$R^{(32)}_{\mathrm{chg}}$	0.319	0.293	0.094	-0.027 [-0.043, -0.015]	$+0.199$ [$+0.167$, $+0.235$]
$R^{(\mathrm{core})}_{\mathrm{chg}}$	0.433	0.390	-0.010	-0.043 [-0.067, -0.020]	$+0.400$ [$+0.319$, $+0.483$]
$\mathrm{MAE}^{(32)}$	2.616	2.630	3.143	$+0.015$ [-0.011, $+0.039$]	-0.513 [-0.559, -0.471]
$\mathrm{MAE}^{(\mathrm{core})}$	2.491	2.518	3.395	$+0.027$ [-0.043, $+0.097$]	-0.877 [-1.003, -0.755]
$\mathrm{MAE}^{(32)}_{\mathrm{adh}}$	1.193	1.3445	1.4033	$+0.151$ [$+0.134$, $+0.169$]	-0.0589 [-0.0946, -0.0190]
$\mathrm{ASSD}^{(32)}$ (mm)	0.3494	0.3487	0.3896	-0.0007 [-0.0012, -0.0002]	-0.0409 [-0.0427, -0.0391]

Per-structure reading of the ablation. The standard diffusion objective lowers hippocampal correlation from 0.388 to 0.228 and ventricular correlation from 0.603 to 0.572; latent control lowers the latter to 0.088. SSIM spans 96.0–96.8%.

Table S12: Volume errors of the Table 4 variants on 669 brain test pairs from 185 subjects. Values are means ± SD across 32 regions or three bilaterally merged core structures; MAE is in percentage points.

Variant	$\mathrm{MAE}^{(32)}$	$\mathrm{MAE}^{(\mathrm{core})}$
GraphMorph (full)	**2.67** ± 1.29	**2.66** ± 0.73
No control map	3.74 ± 1.77	3.34 ± 0.90
Evolver replaced by identity	3.00 ± 1.67	3.47 ± 1.41
Latent control signal	3.44 ± 1.93	3.65 ± 1.88
Standard diffusion objective	3.04 ± 1.68	2.82 ± 0.71

Request supervision. With the forecaster fixed, request-target supervision gives the higher test $R^{(32)}_{\mathrm{chg}}$ on each of three matched renderer seeds (Table S13), and the matched-seed progression errors over 32 regions favour it as well ($\Delta^{(32)}_{\mathrm{RMAE}}$ $+0.014$ to $+0.017$ and $\Delta^{(32)}_{\mathrm{Region}}$ $+0.031$ to $+0.037$, every 95% subject-bootstrap interval above zero), with core differences within ± 0.01. Under the deployed renderer on the default validation canvas, the identity request lowers $R^{(32)}_{\mathrm{chg}}$ from 0.280 to 0.089 and the full-32 decode forecaster's request (Table S2) to 0.214, while reference-request errors remain 1.28–1.52 pp against 1.41 pp for the deployed request: low request error can accompany an inaccurate forecast.

On matched renderer seeds 43–45 and the training canvas, request-target supervision raises validation $R^{(32)}_{\mathrm{chg}}$ for predicted requests ($s = 1$) by $+0.032$ [$+0.023$, $+0.041$] (Table S6). Adherence and forecast errors under both targets are in Table 6.

Table S13: Renderer supervision on 689 brain validation pairs from 186 subjects and 669 test pairs from 185 subjects. Values are means ± sample SD across matched seeds 43–45. Validation uses the default canvas and test the training canvas; adherence MAE is in percentage points.

	Validation, default canvas (689 pairs)			Test, training canvas (669 pairs)	
Target	$R^{(32)}_{\mathrm{chg}}$	$R^{(\mathrm{core})}_{\mathrm{chg}}$	$\mathrm{MAE}^{(32)}_{\mathrm{adh}}$	$R^{(32)}_{\mathrm{chg}}$	$R^{(\mathrm{core})}_{\mathrm{chg}}$
Request-target	**0.282** ± 0.024	0.380 ± 0.037	1.72 ± 0.34	**0.317** ± 0.003	0.418 ± 0.006
Follow-up-target	0.260 ± 0.020	**0.392** ± 0.018	**1.72** ± 0.32	0.285 ± 0.012	**0.428** ± 0.008

B.3 Request scale

On the training canvas for the first 60 validation pairs (12 ADNI subjects), with $s \in \{0, 0.5, 1, 1.5, 2, 3\}$, the deployed checkpoint's pooled response gains are 0.75, 1.11, 1.01 and 0.83

for hippocampus, amygdala, lateral ventricle and temporal horn, responses follow the requested direction monotonically in 75%–100% of pairs, and at $s = 3$ the ventricular request is $+8.7\%$ and the render $+8.9\%$, against median true change $+2.1\%$. The three-scale sweep on all 689 validation pairs (186 subjects; matched renderer seeds 43–45, $s \in \{0, 1, 2.5\}$) retains positive hippocampal and ventricular response slopes against the references, while the pooled amygdala slope is near zero (Table S14); the sweeps measure request response, not forecast accuracy.

Table S14: Three-scale request response on 689 brain validation pairs from 186 subjects, training canvas. Pooled slopes are means ± sample SD across matched seeds 43–45. Brackets give 95% paired subject-bootstrap CIs for request-target minus follow-up-target training.

Structure	Request target	Follow-up target	Request − follow-up [CI]
32 regions	0.037±0.001	0.036±0.002	+0.0011 [+0.0007, +0.0174]
Hippocampus	0.736±0.011	0.765±0.019	−0.0294 [−0.0554, −0.0049]
Amygdala	0.006±0.032	−0.005±0.006	+0.0103 [−0.1015, +0.0152]
Lateral ventricle	0.839±0.011	0.858±0.050	−0.0192 [−0.0323, −0.0060]

B.4 Forecast and renderer variation on the training canvas

Across the tested designs on the training canvas, forecast variation exceeds renderer variation. Across eleven forecasters (five design arms and six reseeds), request-level core correlations span 0.257–0.463 and end-to-end scores span 0.095 (Figure 4, 689 validation pairs). Holding the forecast fixed, seven renderers (four oracle schedules and three reseeds) span 0.014 (Appendix B.5). Renderer variation is larger on the default validation canvas, where seeds 42–45 read $R^{(32)}_{\mathrm{chg}}$ 0.280, 0.308, 0.278 and 0.259 and $R^{(\mathrm{core})}_{\mathrm{chg}}$ 0.417, 0.422, 0.354 and 0.363 (Appendix I.7).

B.5 Oracle schedule

Training the renderer on measured follow-up controls does not improve the deployed forecast. During training the control map is built from the measured follow-up graph with probability p_{oracle} and from the Stage 2 forecast otherwise; the deployed renderer is the p_{oracle}=0 arm, and four schedules were compared with the Stage 2 checkpoint, data, seed and epoch budget fixed, each trained once and evaluated end to end on the same 689 validation pairs and training canvas. The deployed arm has the highest forecast correlations, with core scores spanning 0.421–0.433 across schedules, 32-region correlation 0.319 versus 0.270–0.299 and MAE 2.62 pp. Oracle-trained arms follow measured controls more closely (0.884–0.905 on core and 0.750–0.798 over 32 regions, versus 0.708 and 0.597 for the deployed arm).

B.6 Transfer to a second renderer

The released MambaControl backbone (Yang et al., 2025) shows no response to anatomical-context substitution under the tested settings. On 689 validation pairs, baseline, forecast and measured-follow-up contexts produce identical images at blend weight 0.1; at weight 1, baseline and measured-follow-up contexts remain identical, with sampler variation in forecast-context outputs, and with noise fixed on six pairs even zero, all-0.05 and all-0.95 contexts yield identical latents and images. The six arms span $R^{(\mathrm{core})}_{\mathrm{chg}} = 0.003$–$0.024$ and $R^{(32)}_{\mathrm{chg}} = 0.000$–$0.093$ (Appendix I.6).

B.7 A rendering failure

The rendering failure in Figure S6d is a discrepancy between requested and rendered anatomy on a subject with a dementia diagnosis at every visit.

B.8 The control rollout on one subject

The rendered change follows the request where it is asked: Figure S1 sweeps the requested horizon out to ten years on one test subject observed at eight visits over 7.3 years, with its baseline passed through the deployed system once per horizon, and the displacement field concentrates on the ven-

tricular walls. Training pairs are consecutive visits about seven months apart; the longer-horizon sweep illustrates the control response.

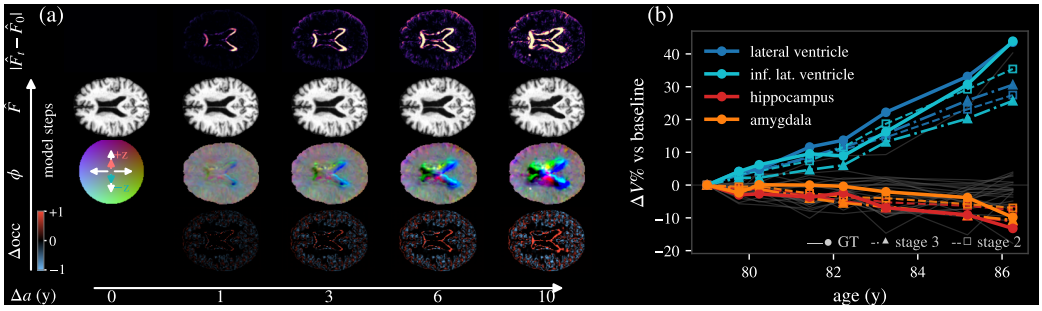

Figure S1: Anatomical control across forecast horizons. (a) Fixed-baseline horizon sweep; (b) composed one-step forecasts along the observed timeline. Control colours use a signed square-root scale.

C Measurement reliability and robustness

C.1 Independent measurement and boundary agreement

Independent segmentation. FastSurfer (Henschel et al., 2020) preserves the 32-region ranking on the 100-pair subset by point estimate (oracle control (F), GraphMorph, Δ-LFM, CardiacAging; Table S15) and narrows GraphMorph's margins. Under SynthSeg the subset reads 0.311 over 32 regions and 0.461 on the core, against 0.314 and 0.422 on the full split; FastSurfer lowers the 32-region reading to 0.234 and raises the MAEs. Oracle control (F) leads the deployed forecast under both segmenters, the paired 32-region lead over Δ-LFM falls from +0.113 [+0.035, +0.183] under SynthSeg to +0.011 [−0.051, +0.075] under FastSurfer (Table S16), and the 32-region MAE advantage over copy-forward reverses (2.66 against 2.80 pp under SynthSeg, 3.45 against 3.20 under FastSurfer).

Table S15: Independent segmentation on 100 brain test pairs, one per subject. SS denotes SynthSeg and FS FastSurfer. Correlation and MAE are means ± SD across 32 regions or three merged core structures. The final column gives FS minus SS with paired subject-bootstrap 95% CIs.

	$R^{(32)}_{\text{chg}}$		$R^{(\text{core})}_{\text{chg}}$		MAE$^{(32)}$ (pp)		$R^{(32)}_{\text{chg}}$: FS − SS
Method	SS	FS	SS	FS	SS	FS	Difference [95% CI]
GraphMorph (ours)	**0.311**±0.152	**0.234**±0.220	**0.461**±0.161	**0.321**±0.214	**2.66**±1.46	**3.45**±2.56	-0.078 [-0.180, +0.013]
GraphMorph, oracle control (F)	0.605±0.202	0.354±0.261	0.697±0.243	0.538±0.333	2.80±1.39	4.12±2.73	-0.251 [-0.315, -0.181]
Δ-LFM	0.198±0.119	0.222±0.193	0.189±0.172	0.217±0.092	5.95±2.91	8.38±5.85	+0.024 [-0.070, +0.098]
CardiacAging	0.129±0.175	0.160±0.212	0.238±0.275	0.276±0.207	3.18±1.61	5.46±4.44	+0.031 [-0.072, +0.122]
Copy-forward ($\hat{F} = M$)	—	—	—	—	2.80±1.69	3.20±2.44	—

Table S16: Paired 32-region correlation differences under SynthSeg (SS) and FastSurfer (FS), on the same 100 brain test pairs, one per subject. Differences are first minus second; brackets give 95% paired subject-bootstrap CIs.

Contrast	$R^{(32)}_{\text{chg}}$: SS	$R^{(32)}_{\text{chg}}$: FS
Oracle control (F) − GraphMorph	+0.293 [+0.238, +0.357]	+0.120 [+0.054, +0.216]
GraphMorph − Δ-LFM	+0.113 [+0.035, +0.183]	+0.011 [−0.051, +0.075]
GraphMorph − CardiacAging	+0.183 [+0.124, +0.240]	+0.074 [−0.001, +0.150]

On the pre-specified independent sample (292 pairs from 84 subjects, compat29), the FastSurfer correlation difference from Δ-LFM is +0.042 [+0.003, +0.086] (Table S17), and +0.042 [+0.006, +0.085] on all 296 pairs with failed segmentations replaced by copy-forward. Post hoc, on the same 292 pairs the difference narrows from +0.102 under SynthSeg to +0.042 under FastSurfer, a change of −0.060 [−0.131, −0.014] that comes from Δ-LFM's correlation rising (+0.044 [−0.005, +0.109]) more than GraphMorph's falls (−0.017 [−0.067, +0.027]). On 581 real test scans the

two segmenters agree on the hippocampus, lateral ventricle and whole brain (CCC 0.903, 0.992 and 0.906) but place the cortex–white-matter boundary differently (volume bias $+21.633\%$ for the left cortex and -16.850% for the left white matter, CCC 0.917 for their union); on rendered images they agree nearly as well as on real scans for GraphMorph (whole-brain CCC 0.887 [0.855, 0.914] against 0.906) and less for Δ-LFM (0.671 [0.611, 0.724]). On the independent 296 pairs their measured-change agreement is $r = 0.522$ [0.484, 0.562] over 32 regions, 0.553 [0.517, 0.589] on compat29 and 0.701 [0.621, 0.761] on the core. Figure S2 shows a typical real scan and the two real scans with the largest amygdala disagreement: over the 200 real scans of the 100-pair subset the FastSurfer-to-SynthSeg amygdala volume ratio has a median of 0.89, and the two scans read 0.66 and 0.67, with the hippocampus, ventricles and thalamus within a few percent on the same scans. On these scans FastSurfer stops short of the amygdala boundary visible in the T1, and the two segmenters' change agreement ($r = 0.553$ on compat29, 0.701 on the core) bounds what a FastSurfer reading can resolve; the retained-subset re-scoring is post hoc and reads -0.000 [-0.062, $+0.063$] (Table 2).

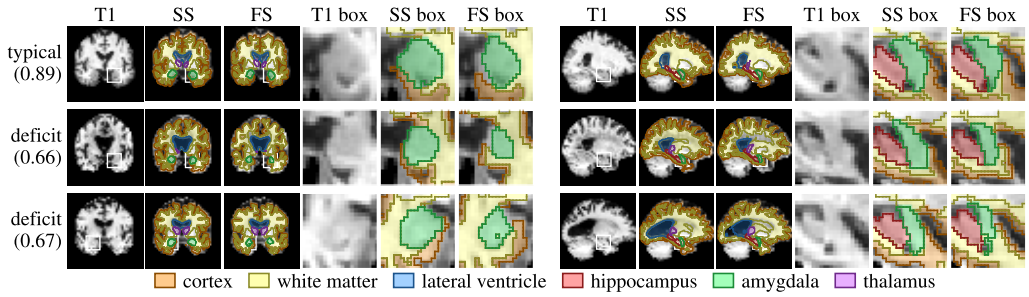

Figure S2: Segmenter disagreement on real test scans. A typical scan (top) and the two scans of the 100-pair subset with the largest FastSurfer amygdala deficit, in coronal and sagittal slices through the amygdala, read by SynthSeg (SS) and FastSurfer (FS). FastSurfer DKT labels are resampled with nearest neighbour and cortical parcels merged into cortex. FastSurfer stops short of the amygdala boundary that SynthSeg follows in both planes, and the two place the cortex–white-matter boundary differently. The hippocampus, ventricles and thalamus coincide.

Table S17: Prespecified independent FastSurfer evaluation on 292 complete brain test pairs from 84 subjects, compatible 29-region map. Differences compare GraphMorph with the named reference. Brackets give 95% paired subject-bootstrap CIs; MAE is in percentage points.

		R_{chg}	MAE
Analysis	Pairs / subjects	GraphMorph − Δ-LFM	GraphMorph − copy-forward
Independent primary / compat29	292 / 84	$+0.042$ [$+0.003$, $+0.086$]	-0.031 [-0.140, $+0.080$]

Boundary agreement. Boundary agreement is dominated by the small anatomical change between visits: at a median interval of 0.59 years, copy-forward is already within 0.35 mm of the follow-up labels, and surface Dice at one voxel exceeds 0.97 for every compared method except BrLP (0.932) and MambaControl backbone (0.952). The deployed render matches copy-forward's 32-region ASSD and has lower ASSD on core6 (paired difference $+0.006$ [$+0.003$, $+0.009$] mm); the other compared methods worsen ASSD, by 0.20 mm for Δ-LFM ($+0.194$ [$+0.186$, $+0.202$]) and from $+0.031$ [$+0.027$, $+0.035$] for CardiacAging to $+0.404$ [$+0.368$, $+0.451$] for BrLP. Oracle control (F) improves ASSD by 0.028 mm over 32 regions and 0.063 mm on core6 (Table S18; metrics in Appendix H.2). Among the design variants, removing the control map costs $+0.065$ [$+0.061$, $+0.069$] mm of 32-region ASSD, the latent control signal $+0.018$ [$+0.016$, $+0.019$], the standard diffusion objective $+0.006$ [$+0.005$, $+0.008$] and an identity evolver $+0.004$ [$+0.003$, $+0.005$]; the deployed render's three seeds read 0.348–0.349 mm and differ from the three follow-up-target seeds by $+0.001$ to $+0.002$ mm.

C.2 The measurement ceiling

The deployed core correlation is below all three core measurement-ceiling estimates; oracle control (F) lies within their range (Figure S3a; estimator and samples in Appendix H.3). The estimate

Table S18: Boundary agreement on 669 brain test pairs from 185 subjects at 1.5 mm resolution. Means pool valid (pair, region) cells; ± is the sample SD of regional means over 32 regions or six core labels. Bold compares forecasters and copy-forward. Empty predicted masks are excluded from the means.

Arm	$\mathrm{ASSD}^{(32)}$ (mm)	$\mathrm{ASSD}^{(\mathrm{core6})}$ (mm)	$\mathrm{SDice}_{1.5}^{(32)}$	$\mathrm{SDice}_{1.5}^{(\mathrm{core6})}$
Copy-forward ($\hat{F} = M$)	0.349±0.053	0.338±0.025	0.990±0.007	0.993±0.001
Δ-LFM	0.543±0.106	0.512±0.061	0.971±0.018	0.979±0.008
CardiacAging	0.379±0.081	0.352±0.033	0.985±0.014	0.991±0.002
GraphMorph (ours)	**0.349**±0.054	**0.331**±0.025	**0.990**±0.007	**0.994**±0.001
GraphMorph, oracle control (F)	0.321±0.073	0.268±0.051	0.990±0.008	0.995±0.002

depends on how residual variation is attributed to segmentation: the default estimator pools residuals over scans, a degrees-of-freedom correction lowers the pooled core figure to 0.570, a quadratic refit in age (127 subjects with four or more visits) raises it to 0.849, and resampling test subjects 2,000 times places the reported estimator between 0.739 and 0.799. The slowly changing brainstem and cerebral cortex have low estimates (0.59 and 0.61), and oracle control (F) exceeds the default estimate in 11 unmerged regions.

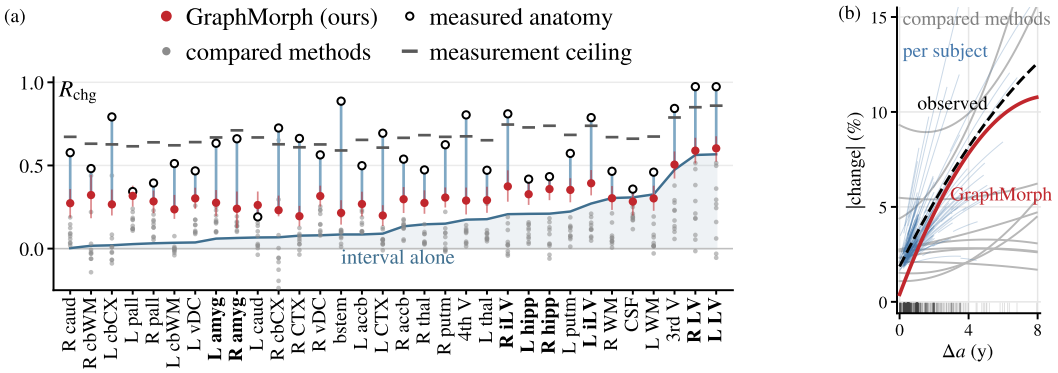

Figure S3: Measurement ceilings and realised change. (a) Unmerged regional change correlations with subject-bootstrap 95% CIs; stems connect deployed and oracle control (F) readings. (b) Realised core-change magnitude by forecast horizon.

The changes themselves are small: over the 669 test pairs the mean absolute change is 2.24% and 2.25% (left and right hippocampus), 3.67% and 4.11% (amygdala) and 5.49% and 5.40% (lateral ventricle), with a median of 2.25% over 32 regions (from 0.89% in the right cerebral cortex to 7.75% in the left temporal horn).

C.3 Sensitivity to amplitude matching

Table S19 compares $s = 1$, validation-fitted scales frozen on test, and test-fitted scales for the reference set, each method keeping the same progression support across scale choices. Brain validation scales use 689 pairs, except Noise-Free Flow, Flow Matching and Diffusion, which use prespecified 186-pair subsets. Brain test uses 669 pairs from 185 subjects, with 590 pairs from 181 subjects retained for core progression. Knee scales use 2,354 retained pairs from a 6,176-pair frame (679 patients), except Diffusion's 270-pair subset (a).

Amplitude matching favours GraphMorph over Δ-LFM on the 32-region paired contrast (GraphMorph minus Δ-LFM): +0.031 [+0.0004, +0.061] at $s = 1$ becomes −0.106 [−0.137, −0.077] under the validation scale and −0.109 [−0.139, −0.079] under the test scale. On the core, differences from Noise-Free Flow, the closest fitted comparator (0.912 [0.864, 0.962] validation-fitted), are −0.069 [−0.127, −0.007], −0.056 [−0.114, +0.001] and −0.064 [−0.121, −0.007], respectively. On the knee joint space, +0.073 [+0.041, +0.104] against Flow Matching and +0.073 [+0.042, +0.104] against Diffusion become −0.026 [−0.061, +0.009] and −0.032 [−0.067, +0.001] under the validation scale. The unrendered knee covariate prediction, measured on the stored-mask ruler and therefore cross-level to the read-back rows, reads 1.396 [1.373, 1.418] at $s = 1$ and 1.154 [1.124, 1.184] validation-fitted.

Table S19: Amplitude matching on brain test pairs (32 regions: 669/185 pairs/subjects; core: 590/181) and 2,850 knee read-back pairs from 776 patients. Entries are pair means on fixed support across scales: $s = 1$ is unscaled, and valid/test name the fitting split. Brackets give 95% subject- or patient-bootstrap CIs. [a]Diffusion uses 270 validation pairs.

(a) Brain, 32 regions

Method	$\Delta^{(32)}_{\text{Region}}$ ($s = 1$)	$\Delta^{(32)}_{\text{Region}}$ (valid)	$\Delta^{(32)}_{\text{Region}}$ (test)	$s_{\text{val}}/s_{\text{test}}$
GraphMorph (ours)	1.200 [1.171, 1.229]	**1.043** [1.016, 1.070]	**1.043** [1.015, 1.070]	2.318 / 2.306
Δ-LFM	**1.170** [1.144, 1.195]	1.149 [1.123, 1.176]	1.152 [1.125, 1.179]	0.747 / 0.810
CardiacAging	1.515 [1.497, 1.533]	1.305 [1.283, 1.326]	1.293 [1.273, 1.313]	2.976 / 3.357
Copy-forward	2.000 [2.000, 2.000]	2.000 [2.000, 2.000]	2.000 [2.000, 2.000]	1.000 / 1.000
Oracle control (F)	0.784 [0.764, 0.805]	0.779 [0.757, 0.800]	0.778 [0.754, 0.801]	0.952 / 0.944

(b) Brain, core

Method	$\Delta^{(\text{core})}_{\text{Region}}$ ($s = 1$)	$\Delta^{(\text{core})}_{\text{Region}}$ (valid)	$\Delta^{(\text{core})}_{\text{Region}}$ (test)	$s_{\text{val}}/s_{\text{test}}$
GraphMorph (ours)	**0.942** [0.890, 0.997]	**0.856** [0.807, 0.907]	**0.855** [0.805, 0.905]	2.318 / 2.306
Δ-LFM	0.997 [0.939, 1.058]	0.988 [0.929, 1.050]	0.987 [0.928, 1.049]	0.747 / 0.810
CardiacAging	1.529 [1.488, 1.566]	1.182 [1.124, 1.237]	1.146 [1.096, 1.197]	2.976 / 3.357
Copy-forward	2.000 [2.000, 2.000]	2.000 [2.000, 2.000]	2.000 [2.000, 2.000]	1.000 / 1.000
Oracle control (F)	0.514 [0.476, 0.552]	0.492 [0.452, 0.533]	0.488 [0.445, 0.532]	0.952 / 0.944

(c) Knee, joint space

Method	$\Delta^{(\text{JS})}_{\text{Region}}$ ($s = 1$)	$\Delta^{(\text{JS})}_{\text{Region}}$ (valid)	$\Delta^{(\text{JS})}_{\text{Region}}$ (test)	$s_{\text{val}}/s_{\text{test}}$
GraphMorph (ours)	1.311 [1.287, 1.335]	**1.194** [1.168, 1.220]	**1.194** [1.168, 1.220]	2.334 / 2.324
Diffusion	**1.238** [1.211, 1.265]	1.226 [1.200, 1.253]	1.226 [1.199, 1.252]	1.360[a] / 1.360
Copy-forward	1.998 [1.998, 1.999]	1.998 [1.998, 1.999]	1.998 [1.998, 1.999]	1.000 / 1.000

The fitted point-estimate ordering is unchanged under independent segmentation and on grain. On the 100-pair subset (SynthSeg and FastSurfer core progression sets of 88 and 78 pairs; FastSurfer uses the SynthSeg validation scale), validation-fitted Δ_{Region} for GraphMorph is 1.127 [1.060, 1.194] on 32 regions and 1.003 [0.867, 1.138] on the core under SynthSeg, against 1.154 [1.093, 1.215] and 1.112 [0.983, 1.243] for Δ-LFM, with paired differences -0.027 [-0.091, $+0.033$] and -0.109 [-0.268, $+0.047$]; under FastSurfer the readings are 1.191 [1.131, 1.251] and 0.928 [0.797, 1.069] against 1.227 [1.168, 1.287] and 1.071 [0.921, 1.225], with differences -0.037 [-0.094, $+0.023$] and -0.143 [-0.284, -0.001]. On grain (1,548 test pairs from 258 samples with each arm's identifiable grains; validation scales from 756 pairs of 126 samples, 755 for Flow Matching), GraphMorph reads 1.130 [1.122, 1.139] at $s = 1$ and 0.756 [0.746, 0.767] validation-fitted ($s_{\text{val}}/s_{\text{test}} = 2.876/2.839$), the lowest fitted forecaster reading, whereas Flow Matching is lowest at $s = 1$ (1.058 [1.038, 1.080]) and reads 1.097 [1.063, 1.137] fitted.

C.4 Calibration of regional change

GraphMorph has calibration slopes closer to one than Δ-LFM and CardiacAging on test and overstates change on validation, predicting more hippocampal and amygdalar atrophy and more ventricular expansion than occurred. On test, hippocampal forecasts have mean bias -0.28 pp, agreement slope $\beta = 0.16$ and calibration slope $b = 0.92$; the ventricular slopes are 0.40 and 0.92 (Table S20). On validation (689 pairs from 186 subjects), bias is -0.492 [-0.651, -0.333] pp for the hippocampus, -0.368 [-0.586, -0.157] for the amygdala and $+0.603$ [$+0.254$, $+0.954$] for the lateral ventricles; the hippocampal calibration slope is below one, 0.736 [0.550, 0.932], and within diagnoses the slopes below one are the CN lateral ventricles (0.748 [0.510, 0.986]), the MCI hippocampus (0.683 [0.431, 0.995]) and the AD amygdala (0.531 [0.214, 0.887]). The validation agreement slopes are 0.170 [0.122, 0.221], 0.155 [0.120, 0.201] and 0.398 [0.312, 0.484] for the three structures. Binned by predicted change on the frozen validation quintiles, hippocampal observation minus prediction is $+0.458$ to $+0.717$ pp in the four most atrophic bins and -0.161 [-0.528, $+0.182$] in the fifth, the amygdala differs from its forecast only in the first bin ($+0.824$ [$+0.199$, $+1.416$]), and the ventricular forecast overshoots in the two most expanding bins (-0.712 [-1.296, -0.113] and -1.290 [-2.303, -0.242]).

C.5 Realised magnitude and its rate

Realised core change saturates with the horizon where Puglisi et al. (2025) report accelerating uncertainty: the quadratic term is -0.139 here (se 0.020, $p<0.001$), whereas the quadratic term for the

Table S20: Regional calibration on 669 brain test pairs from 185 subjects. Bias is predicted-minus-measured change in percentage points; β and b are slopes of prediction on measurement and measurement on prediction, respectively.

Method	$\text{bias}^{(\text{Hipp})}$	$\beta^{(\text{Hipp})}$	$b^{(\text{Hipp})}$	$\text{bias}^{(\text{Amyg})}$	$\beta^{(\text{Amyg})}$	$b^{(\text{Amyg})}$	$\text{bias}^{(\text{LV})}$	$\beta^{(\text{LV})}$	$b^{(\text{LV})}$
GraphMorph (ours)	-0.28	0.16	0.92	+0.12	0.12	0.63	-0.32	0.40	0.92
Δ-LFM	-1.31	0.04	0.01	+2.52	0.27	0.16	+2.58	0.40	0.36
CardiacAging	+1.21	0.07	0.39	+0.73	0.09	0.16	-3.87	0.09	2.92

observed change is not detected (p=0.28) (Figure S3b; mixed-effects specification in Appendix H.4; median interval 0.59 years, 98% inside four years).

Per year of requested interval, GraphMorph realises +1.98% of core change against +1.61% occurred (linear rate; difference +0.37 [−0.08, +0.78]). Adjusting for sex, baseline age, baseline diagnosis or all three gives +2.294, +1.977, +1.955 and +1.940% per year, and the within-subject rate (+2.042) exceeds the between-subject rate (+1.637) by 0.405 points per year (p<0.001). At the same interval, forecasts for AD and MCI baselines realise 0.793 and 0.386 percentage points more absolute core change than for CN (both p<0.001).

Per method (Table S21), GraphMorph is the only forecaster whose magnitude tracks the occurred curve. Δ-LFM overshoots, reaching 20.1% by eight years against the 12.6% that occurred; the other nine (CardiacAging and the eight forecasters not tabulated) rise by under 1.1 points from zero to four years, against the 6.3 points the anatomy moved. Copy-forward realises none of the 3.50% of core volume that moved and wins both PSNR and SSIM (Figure S4).

Table S21: Realised absolute core change on 669 brain test pairs from 185 subjects. Entries are subject mixed-effects estimates in percent; parentheses give Wald p-values for the quadratic coefficient.

Curve	1 year	4 years	8 years	quadratic term (p)
Change that occurred	3.6	8.2	12.6	−0.059 (0.28)
GraphMorph (ours)	2.7	7.8	10.8	-0.139 (1.7×10^{-12})
Δ-LFM	5.1	9.3	20.1	+0.186 (0.0001)
CardiacAging	1.1	2.1	5.4	$+0.071$ (7.1×10^{-8})

Fidelity against realised change. Controlling for the forecast interval weakens the association between image fidelity and realised change. Over the 669 predictions, realised change and requested interval correlate at 0.864, and holding Δa fixed (subject random intercept) moves the PSNR slope on realised change from −0.231 (p<0.001) to −0.102 (p=0.31) and the SSIM slope from −0.0018 to 0.0000 (p=0.99). Across the eleven scored methods and copy-forward, $\rho = -0.50$ on PSNR and −0.80 on SSIM (Figure S4).

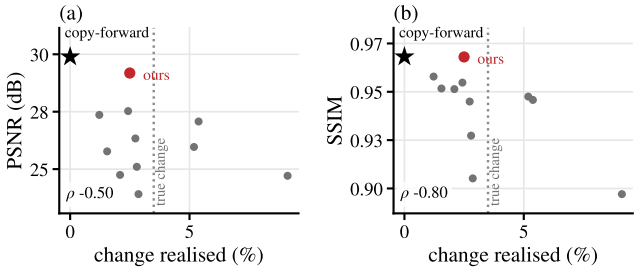

Figure S4: Image fidelity and anatomical change. Follow-up-informed references are excluded.

C.6 STRATA

The correlation lead over Δ-LFM holds in every stratum, and GraphMorph has its lowest change correlations in AD (Table S22). Subject-bootstrap paired differences (2,000 draws) are positive in all twelve stratum–aggregate cells: the 32-region differences span +0.081 to +0.103 with lower limits of at least 0.031, and the core differences +0.104 to +0.364, the smallest in AD with interval [+0.003, +0.226]. Male–female correlation differences are 0.005 across 32 regions and 0.086 on the core.

Strata are baseline diagnosis and recorded sex, scored with the fixed models. Cases with unrecorded sex are all from OASIS-3, with median interval 2.5 years against 0.55 for the rest, so unrecorded sex is not a level. Δ-LFM is the comparator, chosen by the highest test $R^{(32)}_{\mathrm{chg}}$ within each stratum; its MAE ranges from 5.79 to 6.11 pp, and Noise-Free Flow has lower pooled MAE (3.28 pp).

Table S22: Brain test results by baseline diagnosis and recorded sex; n counts pairs (669 overall; 568 with recorded sex). Correlation and MAE show means ± SD across 32 regions or three merged core structures; core change at one year shows the mixed-model mean ± half a 95% Wald interval. Stars mark forecast-minus-observed change intervals excluding zero; daggers mark 95% subject-bootstrap correlation differences from Δ-LFM excluding zero.

Method	$R^{(32)}_{\mathrm{chg}}$	$R^{(\mathrm{core})}_{\mathrm{chg}}$	MAE$^{(32)}$ (pp)	Core change at 1 year (%)
All pairs: $n = 669$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.678*; occurred: $+0.54 \pm 0.14$				
GraphMorph (ours)	**0.314**±0.094 †	**0.422**±0.136 †	**2.67**±1.29	+0.49±0.14
Δ-LFM	0.230±0.125	0.204±0.146	5.90±2.89	+1.86±0.24 *
CN: $n = 186$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.584*; occurred: $+0.42 \pm 0.24$				
GraphMorph (ours)	**0.330**±0.129 †	**0.474**±0.170 †	**2.55**±1.29	+0.40±0.24
Δ-LFM	0.231±0.104	0.289±0.144	5.79±2.94	+1.27±0.41 *
MCI: $n = 318$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.671*; occurred: $+0.72 \pm 0.26$				
GraphMorph (ours)	**0.310**±0.094 †	**0.419**±0.086 †	**2.59**±1.23	+0.45±0.26
Δ-LFM	0.207±0.165	0.055±0.090	5.87±2.84	+2.23±0.43 *
AD: $n = 165$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.666*; occurred: $+0.42 \pm 0.31$				
GraphMorph (ours)	**0.297**±0.124 †	**0.367**±0.192 †	**2.93**±1.47	+0.52±0.31
Δ-LFM	0.216±0.109	0.263±0.159	6.11±3.08	+1.89±0.48 *
Female: $n = 315$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.588*; occurred: $+0.51 \pm 0.23$				
GraphMorph (ours)	**0.301**±0.100 †	**0.383**±0.089 †	**2.46**±1.17	+0.35±0.23
Δ-LFM	0.208±0.170	0.078±0.100	5.80±2.74	+2.77±0.41 *
Male: $n = 253$; ceiling $R^{(32)}_{\mathrm{chg}}$: *0.648*; occurred: $+0.71 \pm 0.23$				
GraphMorph (ours)	**0.306**±0.116 †	**0.469**±0.126 †	**2.69**±1.32	+0.48±0.23
Δ-LFM	0.218±0.092	0.172±0.099	5.83±3.02	+0.99±0.37

The signed level at one year. At a common one-year interval, realised and measured signed core change differ by -0.05 pp on validation ($p = 0.58$) and -0.04 on test ($p = 0.61$) in the pooled sample, by $+0.07$ and -0.02 pp in controls, and by -0.28 pp on test ($p = 0.075$) and -0.13 on validation ($p = 0.35$) in MCI. The common interval compares strata with different follow-up schedules (median 0.52 years in MCI, about 0.95 elsewhere; protocol in Appendix H.4).

Stratum contrasts from the pooled interaction model do not replicate across splits: diagnosis contrasts have $p>0.35$ on both splits and the sex contrast $p=0.45$ on test, the female gap is -0.31 ($p=0.047$) on validation but -0.16 ($p=0.26$) on test, and the AD stratum, mostly OASIS-3 with long intervals and 24 ADNI subjects on test, is sensitive to the analysis frame. Δ-LFM overestimates one-year core change (Table S22).

C.7 Sensitivity to image-grid origin

A different header origin in the intermediate 1 mm images shifts the placement of 38 subjects (205 scans; 37 on test, one on validation, none in training) by 8 mm along each axis on the 1.5 mm grid (Appendix J); no pair spans both placements. Dropping them moves every ADNI PSNR cell of the cohort benchmark (Appendix D) by less than 0.12 dB except BrLP's, which rises from 24.51 to 25.32, past Diffusion and Pix2Pix. For the same predictions, GraphMorph's change aggregates read 0.324 and 0.435 on the 507 majority-placement pairs and 0.300 and 0.430 on the 162 shifted pairs, against 0.314 and 0.422 on all 669; the compared methods keep their ordering in both groups, and both autoencoder ceilings are flat across placements.

C.8 Sensitivity to segmentation outliers

Excluding the subject with the flagged visit changes the 32-region correlation from 0.314 to 0.315. Checks cover 27,328 truth-side test volumes for zero or missing values and 21,408 consecutive-pair regional changes for halving or doubling and for reversal at the next visit. No volume is zero or missing. The flagged visit, a relabelled right cerebellar cortex, is retained with its two pairs;

excluding the subject changes correlations in four cerebellar regions by up to 0.35 (Δ-LFM's right cerebellar cortex) and other rows by at most 0.025. Flagged validation visits were retained in the 689-pair set.

D BRAIN MRI: COMPLETE BENCHMARK RESULTS

D.1 REFERENCES AND COHORT COMPARISONS

The pooled ranking holds within each cohort (Table S25): GraphMorph leads the reference set on core-change correlation and core MAE, and copy-forward on PSNR, in ADNI, AIBL and OASIS-3 alike.

Table S23: Brain test results on 669 pairs from 185 subjects. Correlation and MAE are means ± SD across 32 regions; MAE is in percentage points. $^{*}p < 0.05$, $^{***}p < 0.001$ versus elapsed interval (test fit) (uncorrected).

Method	$R^{(32)}_{\text{chg}}$ ↑	$\text{MAE}^{(32)}$ ↓
DiffuseMorph	0.074±0.060	4.31±1.95
BrLP	0.069±0.062	9.51±4.62
MambaControl backbone	0.054±0.084	7.79±4.41
Δ-LFM	0.230±0.125*	5.90±2.89
CardiacAging	0.102±0.120	3.32±1.62
ImageFlowNet	0.122±0.131	3.75±1.81
Pix2Pix	0.122±0.090	4.75±1.82
Noise-Free Flow	0.087±0.099	3.28±1.49
Flow Matching	0.122±0.092	4.03±1.61
Diffusion	0.092±0.104	4.38±1.96
GraphMorph (ours)	**0.314**±0.094***	**2.67**±1.29
Copy-forward	—	2.94±1.70
Elapsed interval (test fit)	0.165±0.150	—
Elapsed interval (validation-selected)	0.174±0.148	—
All-region tabular	0.369±0.100	—

Table S24: Follow-up-informed brain reference readings on 669 test pairs from 185 subjects. Correlation and MAE are means ± SD across three core structures; image metrics use pair SD. MAE is in percentage points. (F) uses the measured follow-up; other references appear in Table 1.

Reading	$R^{(\text{core})}_{\text{chg}}$ ↑	$\text{MAE}^{(\text{core})}$ ↓	PSNR↑	SSIM %↑
DiffuseMorph, registration reading (F)	0.540 ± 0.158	3.87 ± 0.34	27.39 ± 1.56	95.69 ± 0.99
MAISI ceiling: true follow-up encoded–decoded (F)	—	—	29.87 ± 1.61	97.44 ± 0.49
Δ-LFM ceiling: true follow-up encoded–decoded (F)	—	—	32.51 ± 1.34	98.47 ± 0.20

Table S25: Brain test results by cohort (669 pairs from 185 subjects; cohort sizes shown). Core correlation and MAE are means ± SD across three core structures; PSNR/SSIM use pair SD. Bold marks the best of all eleven benchmark methods, including those omitted here (Table S23).

Method	$R^{(\text{core})}_{\text{chg}}$ ↑	$\text{MAE}^{(\text{core})}$ ↓	PSNR↑	SSIM %↑
ADNI: 536 pairs				
Δ-LFM	0.111±0.078	4.64±0.62	26.04±1.57	95.59±1.53
CardiacAging	0.224±0.116	3.31±0.91	27.53±2.32	96.04±2.14
GraphMorph (ours)	**0.430**±0.113	**2.48**±0.61	**29.44**±2.90	**97.05**±2.08
Copy-forward	—	—	30.21±3.16	97.06±2.11
AIBL: 32 pairs				
Δ-LFM	0.137±0.160	4.10±0.45	26.03±1.42	95.82±0.99
CardiacAging	0.209±0.105	3.67±1.35	28.12±1.99	96.62±1.16
GraphMorph (ours)	0.369±0.077	**2.20**±0.54	**30.50**±2.61	**97.75**±1.12
Copy-forward	—	—	31.41±2.82	97.75±1.15
OASIS-3: 101 pairs				
Δ-LFM	0.289±0.156	6.06±2.78	25.55±1.38	89.98±7.81
CardiacAging	0.268±0.201	4.43±2.39	26.20±1.52	94.16±1.78
GraphMorph (ours)	**0.373**±0.171	**3.73**±1.75	**27.40**±1.86	**95.20**±1.78
Copy-forward	—	—	27.74±1.92	95.18±1.82

Table S26: Paired brain test contrasts on 669 pairs from 185 subjects, for averages over 32 regions or three core structures. Subtraction follows the row labels. The two 95% CI columns resample subjects and pairs, respectively; p is one-sided. Panel (b) reports subject-CI midpoints.

(a) Forecast and rendering contrasts

Contrast	Quantity	Difference	95% CI, subjects	95% CI, pairs	p
GraphMorph − Δ-LFM	$R^{(32)}_{\text{chg}}$	+0.084	[+0.052, +0.133]	[+0.051, +0.131]	< 0.001
GraphMorph − CardiacAging	$R^{(\text{core})}_{\text{chg}}$	+0.153	[+0.102, +0.208]	[+0.093, +0.215]	< 0.001
GraphMorph − Δ-LFM	$R^{(\text{core})}_{\text{chg}}$	+0.218	[+0.150, +0.293]	[+0.147, +0.292]	< 0.001
GraphMorph − CardiacAging	$R^{(32)}_{\text{chg}}$	+0.212	[+0.189, +0.235]	[+0.181, +0.241]	< 0.001
Copy-forward − GraphMorph (MAE)	$\text{MAE}^{(\text{core})}$	+0.844	[+0.696, +1.003]	[+0.706, +0.988]	< 0.001
Copy-forward − GraphMorph (MAE)	$\text{MAE}^{(32)}$	+0.278	[+0.218, +0.339]	[+0.222, +0.336]	< 0.001
Full − evolver replaced by identity	$R^{(\text{core})}_{\text{chg}}$	+0.437	[+0.361, +0.518]	[+0.367, +0.507]	< 0.001
Full − evolver replaced by identity	$R^{(32)}_{\text{chg}}$	+0.253	[+0.225, +0.290]	[+0.222, +0.287]	< 0.001
Full − zero control	$R^{(\text{core})}_{\text{chg}}$	+0.380	[+0.314, +0.446]	[+0.312, +0.445]	< 0.001
Full − zero control	$R^{(32)}_{\text{chg}}$	+0.253	[+0.225, +0.280]	[+0.220, +0.282]	< 0.001
Full − standard diffusion objective	$R^{(\text{core})}_{\text{chg}}$	+0.100	[+0.048, +0.152]	[+0.048, +0.153]	< 0.001
Full − standard diffusion objective	$R^{(32)}_{\text{chg}}$	+0.136	[+0.116, +0.156]	[+0.111, +0.159]	< 0.001
Full − latent control signal	$R^{(\text{core})}_{\text{chg}}$	+0.315	[+0.254, +0.374]	[+0.247, +0.381]	< 0.001
Full − latent control signal	$R^{(32)}_{\text{chg}}$	+0.249	[+0.227, +0.274]	[+0.221, +0.276]	< 0.001

(b) Method minus elapsed-interval reference (test fit)

Contrast	Quantity	Subject CI midpoint	95% CI, subjects	95% CI, pairs	p
GraphMorph (ours) − interval reference	$R^{(32)}_{\text{chg}}$	+0.146	[+0.123, +0.170]	[+0.117, +0.172]	< 0.001
GraphMorph (ours) − interval reference	$R^{(\text{core})}_{\text{chg}}$	+0.126	[+0.079, +0.173]	[+0.078, +0.174]	< 0.001
Δ-LFM − interval reference	$R^{(32)}_{\text{chg}}$	+0.053	[+0.005, +0.102]	[+0.002, +0.107]	0.017
Δ-LFM − interval reference	$R^{(\text{core})}_{\text{chg}}$	−0.095	[-0.161, -0.029]	[-0.159, -0.029]	0.999
CardiacAging − interval reference	$R^{(32)}_{\text{chg}}$	−0.066	[-0.094, -0.038]	[-0.097, -0.037]	1.000
CardiacAging − interval reference	$R^{(\text{core})}_{\text{chg}}$	−0.027	[-0.089, +0.035]	[-0.094, +0.038]	0.806

D.2 Regional variation

GraphMorph exceeds its elapsed-interval reference in every region except cerebral white matter and CSF; Figure S3 shows the per-region rungs for the deployed model, oracle control (F) and measurement ceiling.

For the bilaterally merged hippocampus, amygdala, lateral ventricle and temporal horn, GraphMorph reaches 0.388, 0.274, 0.603 and 0.420, against elapsed-interval references 0.237, 0.080, 0.570 and 0.262. The first, second and fourth comparisons have $p<0.001$; the ventricular comparison has $p = 0.13$. The strongest compared forecasters reach 0.169 (CardiacAging), 0.206^{*} (Δ-LFM), 0.540 (ImageFlowNet) and 0.199 (Flow Matching), respectively.

D.3 Change magnitude and paired comparisons

Paired subject-bootstrap intervals of the progression errors favour GraphMorph against every compared forecaster on both aggregates; the narrowest margins are against Noise-Free Flow ($\Delta^{(32)}_{\text{RMAE}}$ $+0.121$ [$+0.102$, $+0.142$]; $\Delta^{(\text{core})}_{\text{RMAE}}$ $+0.070$ [$+0.024$, $+0.117$]).

Table S26(a,b) gives the subject- and pair-bootstrap intervals of the main comparisons. Independent re-segmentation changes the MAE ranking (Appendix C.1).

Image fidelity across scoring masks. Copy-forward leads GraphMorph on 97.2% of pairs with whole-volume PSNR and 97.6% with masked PSNR, no forecaster reaches its copy-forward floor under either score, and the method ranking is the same under both except that Pix2Pix and Diffusion exchange rank.

D.4 Qualitative examples

Figure S5 shows longitudinal forecasts across the three datasets.

On a random test subject (Figure S6b), deformation-based methods keep smaller tissue residuals than appearance-generating methods. The figure also extends the main comparison to all methods (a) and shows forecasting and rendering errors (c,d; Appendices A.5 and B.7).

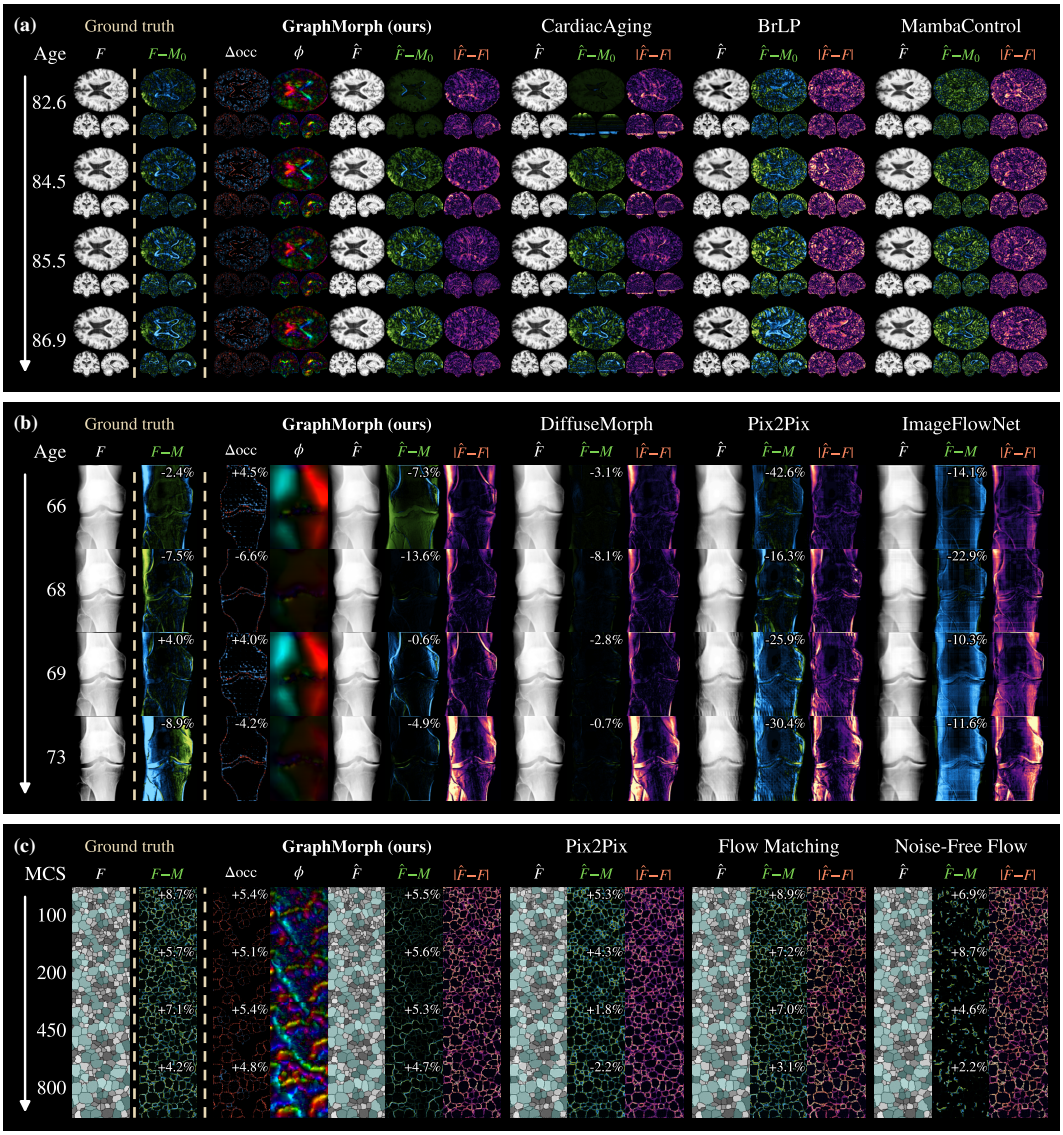

Figure S5: Longitudinal forecasts across three datasets. (a) Brain MRI. Predictions use the preceding visit. Cumulative changes are relative to the first visit (M_0), while Δocc and ϕ describe each row's interval. (b) Knee radiographs. The last prediction uses the unshown age-71 visit. (c) A validation grain microstructure. Changes in (b,c) are relative to each row's input. Percentages show joint-space and median grain-area changes, respectively. Each map type shares a scale within its panel.

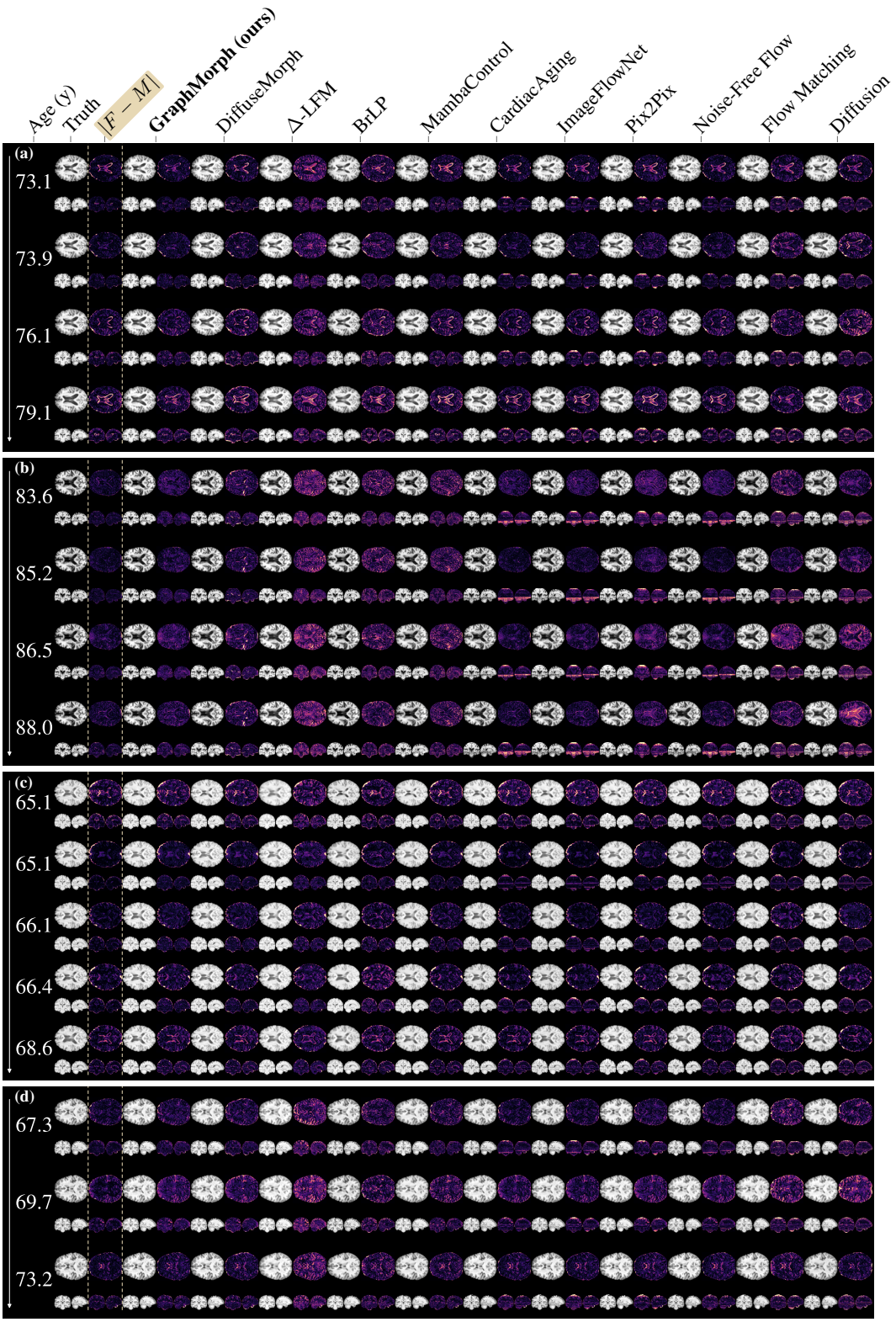

Figure S6: Brain MRI method comparisons. (a) The subject in Figure 2. (b) A random test subject. (c) Forecasting error, including a repeat scan one day after the first follow-up. (d) Rendering discrepancy. Predictions use the preceding visit. Method blocks show predictions and absolute follow-up errors, with a shared magma scale within each panel.

E KNEE RADIOGRAPHS

The knee task forecasts femur, tibia and joint-space anatomy on OAI radiographs (Appendix H.6).

E.1 BENCHMARK AND QUALITY CONTROL

Table S27: Knee joint-space and image prediction on 7,517 strict-QC test pairs from 835 patients. R_{chg}/MAE use stored-mask truth; progression errors use read-back truth on 2,850 pairs. The tabular reference retains the stored-mask baseline. Correlation ± is pair-bootstrap SD; other ± values are pair SD. (F) uses measured follow-up anatomy; stars and daggers follow Table 3.

Method	$R^{(\mathrm{JS})}_{\mathrm{chg}}$ ↑	$\Delta^{(\mathrm{JS})}_{\mathrm{RMAE}}$ ↓	$\Delta^{(\mathrm{JS})}_{\mathrm{Region}}$ ↓	MAE$^{(\mathrm{JS})}$ (pp)↓	PSNR↑	SSIM %↑
GraphMorph (ours)	**0.230*****±0.016	**0.899**±0.533	**1.194**±0.741	**9.53**±8.87	21.95±6.35	63.57±25.49
Pix2Pix	0.093**±0.022	1.174†±1.149	1.243†±0.731	13.53±16.25	19.20±5.34	**65.57**±26.05
Diffusion	0.090**±0.022	1.072†±0.909	1.226±0.741	12.47±13.78	13.12±4.00	33.14±19.86
Noise-Free Flow	0.087**±0.018	0.931†±0.419	1.271†±0.716	9.63±9.13	22.16±7.26	63.60±26.98
Flow Matching	0.119**±0.024	1.031†±0.969	1.220±0.744	11.72±11.15	21.70±5.67	65.05±25.34
ImageFlowNet-GraphODE	0.104**±0.023	1.110†±0.861	1.262†±0.744	12.54±11.13	22.17±5.02	65.36±22.25
DiffuseMorph	0.069±0.017	0.989†±0.532	1.337†±0.704	10.03±9.59	**22.89**±6.89	64.06±25.22
Copy-forward	0.022±0.023	1.000†±0.005	1.998†±0.013	8.91±8.72	23.38±7.24	64.06±25.67
Elapsed interval	0.038±0.011	—	—	8.92±8.70	—	—
Tabular reference	0.474±0.019	0.902±0.441	1.154†±0.739	8.01±7.49	—	—
Read-back ceiling (F)	0.650±0.011	—	—	6.89±6.76	—	—
Oracle control (F)	0.755±0.010	0.562†±0.525	0.678†±0.625	5.77±6.20	23.79±7.29	65.33±25.55
DiffuseMorph, registration (F)	0.377±0.024	0.600†±0.671	0.688†±0.629	8.85±10.23	24.69±7.87	70.62±26.30

GraphMorph leads the evaluated image forecasters on joint-space change correlation (Table S27). Its margin over Flow Matching is 0.111 [0.063, 0.161], and the 95% patient-bootstrap intervals are positive against all six competing image forecasters (Table S29). It also has the lowest Δ_{RMAE} on segmenter-measured joint-space change among image forecasters; over three regions, Noise-Free Flow and DiffuseMorph are lower (1.136 and 1.254 against 1.324). Near-uniform visit spacing leaves elapsed time weakly predictive (0.025 over three regions, 0.038 on joint space).

QC filtering changes the readings little: strict QC retains 7,517 of 8,217 test pairs, read-back joint-space correlation is 0.650 on retained pairs and 0.490 on the 700 excluded, and without filtering GraphMorph and Flow Matching read 0.223 and 0.113 against 0.230 and 0.119.

Flagged joint-space pairs are retained on both splits. Truth-side checks find no zero areas on any split and no bone halving, doubling or reversal on test; the remaining flags affect 113 retained pairs (1.5%) carrying 19% of measured change variance, an independent reading confirms their changes ($r = 0.766$; matching signs on 100 of the 113), and excluding them lowers GraphMorph, Flow Matching and read-back correlations to 0.211, 0.111 and 0.641 with the ranking unchanged.

Table S28: Knee input-control comparison on 7,517 test pairs from 835 patients. Joint-space change uses stored-mask truth; MAE is in percentage points.

Method	$R^{(\mathrm{JS})}_{\mathrm{chg}}$	MAE$^{(\mathrm{JS})}$
GraphMorph	0.230	9.53
Pix2Pix	0.093	13.53

Table S29: Paired knee joint-space correlation differences on 7,517 test pairs from 835 patients: GraphMorph minus each comparator. Brackets give 95% patient-bootstrap CIs, retaining both knees and all visits. The unrendered tabular comparison is cross-level; (F) uses measured follow-up information.

Comparison	$R^{(\mathrm{JS})}_{\mathrm{chg}}$ difference [95% CI]
GraphMorph − DiffuseMorph	+0.161 [+0.135, +0.188]
GraphMorph − ImageFlowNet-GraphODE	+0.126 [+0.086, +0.166]
GraphMorph − Pix2Pix	+0.136 [+0.092, +0.176]
GraphMorph − Noise-Free Flow	+0.143 [+0.103, +0.180]
GraphMorph − Flow Matching	+0.111 [+0.063, +0.161]
GraphMorph − Diffusion	+0.140 [+0.092, +0.179]
GraphMorph − Copy-forward	+0.208 [+0.158, +0.268]
GraphMorph − Tabular reference (unrendered; cross-level)	−0.244 [−0.287, −0.202]
GraphMorph − Oracle control (F)	−0.525 [−0.558, −0.493]
GraphMorph − DiffuseMorph, registration (F)	−0.147 [−0.212, −0.085]

E.2 WHAT THE FORECAST CARRIES AND WHERE IT ERRS

GraphMorph's regional correlation lead holds on joint space and tibia; on femur, four methods exceed it (0.138–0.154 against 0.133). Measured change is small on the bones (SD 3.0% femur, 3.1% tibia) against 12.5% on joint space.

GraphMorph's joint-space forecasts lead image forecasters in ranking and are compressed in amplitude: their spread is 6.6 pp against 12.5 pp measured, a derived calibration slope of 0.43, against 0.81 for the follow-up read-back and 1.01 for the unrendered tabular reference on the stored-mask ruler (derived slopes, without intercept or interval; Appendix H.2). Over the three regions its MAE exceeds copy-forward's (5.78 against 4.43 pp), with the excess on femur and tibia; among predictions without follow-up information, only the tabular reference has lower mean error than copy-forward, on joint space (Table S27) as over the three regions.

The tabular reference's lead does not rest on the disease grade. Without the knee's visit-00 grade its joint-space correlation is 0.464 [0.428, 0.504] against 0.474 [0.438, 0.515] with it, and 0.461 [0.423, 0.505] after also removing body-mass index, sex and age (knee-bootstrap intervals); the grade accounts for +0.010 [+0.005, +0.016] (patient-bootstrap interval). These grade-ablation results concern the tabular reference; GraphMorph uses the grade.

Bias before and after rendering. The decoded requests are already biased, and rendering adds a further offset. On 6,608 registration-QC validation pairs, request correlations reach 29%, 68% and 35% of the femur, joint-space and tibia tabular references (Table S30); femur and tibia biases are −3.59 and −1.71 points against measured mean changes of 0.11% and 0.23%, and joint-space bias is +1.25 points against 0.42%. On the stricter 6,176-pair validation frame, request, render and tabular correlations are 0.295, 0.214 and 0.431, and delivered change follows requested change at $r = 0.54$ and slope 0.67 with a mean of −2.4% against +1.6% requested and +0.4% measured. The oracle-control (F) render has the same offset direction, −1.5% on average.

Table S30: Knee representation and forecaster design, reported separately for femur, joint space and tibia. Stage 1 uses 8,075 validation masks (mean Dice ± mask SD); Stage 2 uses 6,608 registration-QC validation pairs (correlation ± pair-bootstrap SD). The six-reseed row instead reports mean ± SD across seeds 44–49.

Arm	Femur	Joint space	Tibia
(a) Stage 1: follow-up round trips, Dice ↑			
Deployed: 2-D, three regions, 16×16 latent grid	0.9995±0.0005	0.9986±0.0022	0.9995±0.0008
(b) Stage 2: decoded requests, R_{chg} ↑			
Tabular reference	0.419±0.013	0.446±0.017	0.448±0.017
Elapsed interval alone	0.003±0.008	0.020±0.012	0.006±0.008
Baseline: the brain's objective (λ_c=1)	0.093±0.013	0.247±0.017	0.127±0.013
Deployed: λ_c=0 (chosen on the JS column)	0.120±0.013	0.305±0.014	0.156±0.015
Deployed − per-step Δt scaling	0.068±0.013	0.313±0.020	0.200±0.018
Deployed − previous-interval rate	0.096±0.013	0.203±0.019	0.167±0.016
Deployed, six reseeds (44–49)	0.118±0.014	0.281±0.021	0.148±0.020

The Stage 2 selection. Dropping the brain's correlation-loss term, selected on validation, is the deployed knee setting: it reduces bone overprediction (femur/tibia spreads fall from 7.4/5.2 to 3.6/2.0 points, against measured 3.1/3.4) and raises joint-space correlation from 0.247 to 0.305. Removing historical rates lowers joint-space correlation to 0.203; removing per-step Δt scaling leaves it at 0.313 but lowers femur correlation from 0.120 to 0.068. Six Stage 2 reseeds give joint-space correlation 0.281 ± 0.021 (range 0.252–0.311), with bone bias in every run; the prespecified deployed seed 42 is at the range's upper end. Epoch budget and checkpoint selection: Appendix H.6.

E.3 RENDERING AND SUBSTITUTION

Oracle control (F), the anatomy measured on the true follow-up rendered through the same frozen checkpoint, raises the joint-space reading from 0.230 to 0.755 on the 7,517 strict-QC test pairs and from 0.214 to 0.760 on the 6,176 strict-QC validation pairs, above the read-back ceilings of 0.650 and 0.608 (Table S31; paired bootstrap of the test difference, $[0.083, 0.129]$). With the measured request the rendered change has slope 0.65 on validation, so slope and offset also carry rendering error. Scored as an image on the same test pairs, the oracle-control render reads PSNR 23.79, SSIM

65.3, bone Dice 98.1 and a joint-space width error of 0.50 pixels against 21.95, 63.57, 96.6 and 0.67 deployed, clearing the copy-forward floor on all four. The selected knee in Figure S5b loses 14.6% of its joint-space area over eight years; its request has the sign of the truth in three of its four rows.

The control path. Amplifying the request hurts, and control-path variants move joint-space correlation little. On the 6,176 strict-QC validation pairs, removing request information lowers joint-space correlation (Table S31); scaling the request by 1.5 or 2.0 raises errors in all three regions and reads 0.209 and 0.138 against 0.214. Latent control lowers it (0.188); the standard objective raises it slightly (0.218) and reduces femur error (2.31 against 4.63 pp). The oracle schedules (Appendix B.5) read 0.193, 0.208 and 0.211 for p_{oracle} of 1, 0.5 and 1→0, and always-oracle training also reduces femur error (2.98 pp). These single-run comparisons use the deployed seed and schedule.

Request supervision. Follow-up-target training trades adherence for accuracy on the knee: across three seeds per training target it raises joint-space correlation on the same 6,176 validation pairs, with the same ordering on test (Table S31), but worsens femur adherence. Identity requests lower joint-space correlation under both objectives (0.083 and 0.135 at seed 42); reseeded requests stay within 0.02 of the deployed request's reading.

Through the frozen renderer, which was trained on graph requests, tabular-derived masks raise joint-space correlation over full-graph requests by 0.063 [0.041, 0.084] and lower MAE by 0.44 [0.33, 0.54] pp; self-only requests read lower correlation (−0.040 [−0.057, −0.022]) and higher MAE (+0.92 [+0.81, +1.03] pp) than full-graph requests. With matched renderer seeds, request-target training reduces joint-space adherence error (5.16 against 5.76 pp) but increases forecast MAE (9.20 against 9.06 pp); the joint-space monotone fraction is 0.085 (request-target) and 0.099 (follow-up-target).

Table S31: Knee joint-space renderer supervision and request substitutions: 6,176 strict-QC validation pairs from 679 patients and 7,517 test pairs from 835 patients. Correlation ± is pair-bootstrap SD; MAE ± is pair SD; brackets are 95% pair-bootstrap CIs. Each seed is reported separately. Request substitutions retain the renderer; (F) uses measured follow-up anatomy.

	Validation		Test
Renderer training target / request	$R^{(\text{JS})}_{\text{chg}}$ ↑	MAE$^{(\text{JS})}$ (pp)↓	$R^{(\text{JS})}_{\text{chg}}$ ↑
Request-target, deployed (seed 42)	0.214±0.016	9.26±8.39	0.230 [0.197, 0.262]
Identity request (no change)	0.083±0.021	9.36±8.60	—
Zero control (at inference)	0.062±0.019	10.28±8.78	—
Oracle control (F): substituted follow-up anatomy	0.760±0.009	5.57±5.75	0.755±0.010
Request-target, seed 43	0.213±0.016		0.223 [0.191, 0.255]
Request-target, seed 44	0.218±0.016	—	0.222 [0.191, 0.254]
Follow-up-target, seed 42	0.261±0.018	—	0.272 [0.243, 0.301]
Follow-up-target, seed 43	0.244±0.018		0.266 [0.236, 0.296]
Follow-up-target, seed 44	0.236±0.016	—	0.269 [0.237, 0.301]
Read-back ceiling (F)	0.608±0.022	6.74±7.04	0.650±0.011

E.4 Image fidelity and the follow-up channel

Image-fidelity and anatomical-change rankings differ on the knee too, by a smaller margin than on the brain: copy-forward beats GraphMorph on PSNR in 64.4% of pairs and on SSIM in 56.0%, against 97.2% on the brain (Table S27; Appendix H.2). With its follow-up input restored, the DiffuseMorph checkpoint reads 0.377 on test joint space, a registration reading above every forecast and below the tabular reference's 0.474 (margin −0.097 [−0.157, −0.036], paired bootstrap); with the follow-up withheld by the inference flag alone, 0.069.

E.5 Two ceilings

The two ceilings bound different readings. The read-back ceiling (Table S27, 0.650 on test) measures agreement between segmentations of a real radiograph and its stored mask, and a render driven by that mask can exceed it. The variance-based estimator of Appendix C.2, on the knee a per-knee fit of log area on visit time, charges every departure from a smooth trajectory: refitted over the 1,611 test knees with three or more strict-QC visits it gives 0.591 on the joint space (knee resampling

0.577–0.606) and 0.574 over the three regions, and its quadratic and degrees-of-freedom variants span 0.061–0.730; consecutive changes of one knee are anticorrelated at −0.43 on the joint space, so much of each measured change reverts at the next visit. The deployed 0.230 sits below every variant but the degrees-of-freedom-corrected linear one, and the oracle-control render's 0.755 clears them all. The variance-based ceiling bounds forecasts whose information is independent of that visit-level fluctuation; the oracle-control render, handed the measured follow-up, and the tabular reference, carrying the previous-interval rate, are not.

F GRAIN-GROWTH EXTENSION

Retrained on simulated two-dimensional Potts-model grain growth, the pipeline leads every image forecaster on change correlation and trails the tabular reference. The graph tracks boundary evolution in the 32 largest baseline grains that survive to the final snapshot. Appendix J.3 describes the splits by microstructure and the labeller, the fixed rule that reads every image.

Test comparison. GraphMorph has the lowest Δ_{RMAE} and amplitude-matched Δ_{Region} of the compared forecasts on jointly identifiable grains, under read-back (0.745 and 0.755 against 0.955 and 1.095 for Pix2Pix, the nearest competitor) and exact-label measurement (0.741 and 0.773 against 0.957 and 1.098; Appendix C.3); only GraphMorph and ImageFlowNet-GraphODE receive elapsed time. The tabular reference is a gradient-boosted regressor fitted on the training pairs' exact areas from baseline area, size rank, side count (the von Neumann–Mullins covariate (Mullins, 1956)), elapsed interval, snapshot time, grain count and previous-interval rate.

Rendering and one misread cell account for the gap between the decoder-space Pearson correlation of 0.444 [0.429, 0.458] and the rendered 0.247 [0.151, 0.411] (Spearman 0.342; Table S32): one of 49,456 cells is a 99-pixel grain misread as its invader after shrinking (+1,757% versus true −24%), and its post hoc exclusion gives Pearson 0.397, or 0.89 of the decoder-space correlation.

Table S32: Grain forecasts on a 1,548-pair test frame. Pearson R_{chg} and Spearman ρ pool recognised percentage-change cells; n counts these cells. Brackets give 95% pair-bootstrap CIs. Image metrics show mean ± CI half-width, with SSIM and Dice in percent. Stars compare with the elapsed-interval reference ($^{***}p < 0.001$).

Method	R_{chg} [95% CI]↑	ρ ↑	Cells n	PSNR↑	SSIM↑	Boundary Dice↑
GraphMorph (ours)	**0.247** [0.151, 0.411]***	**0.342**	49,456	12.67±0.02	32.86±0.09	65.8±0.2
Pix2Pix	0.143 [0.117, 0.169]***	0.301	48,768	13.17±0.01	**34.08**±0.09	**68.7**±0.1
Diffusion	-0.116 [-0.146, -0.086]	0.002	11,181	10.45±0.03	19.05±0.43	56.5±0.3
Noise-Free Flow	0.081 [0.061, 0.101]	0.288	48,388	12.22±0.02	29.49±0.08	63.3±0.1
Flow Matching	0.093 [0.067, 0.118]	0.303	42,270	11.95±0.02	31.46±0.09	61.6±0.2
ImageFlowNet-GraphODE	-0.141 [-0.161, -0.120]	0.011	29,295	**14.51**±0.02	30.71±0.08	52.9±0.4
DiffuseMorph	-0.004 [-0.015, 0.007]	0.003	49,459	12.20±0.02	30.79±0.09	63.6±0.2
Copy-forward ($\hat{F} = M$)	—	—	49,459	11.90±0.02	28.31±0.08	62.5±0.2
Tabular reference	0.608 [0.599, 0.617]	0.552	49,459	—	—	—
Read-back ceiling (F)	0.993 [0.993, 0.993]	0.991	49,459	—	—	—

Stage by stage on validation. Stage 1 loses almost nothing and Stage 3 keeps 0.92 of the Stage 2 correlation (Table S33; no Stage 2 configuration was selected on validation). On 756 baseline masks (24,192 mask–grain cells), the 32-grain, 16×16 round trip has relative area MAE **0.218**±0.200% and bias **-0.002**±0.296%.

Forecast correlation is 0.447 before rendering and 0.411 after it. Removing the control map at inference gives −0.047, and the DiffuseMorph backbone alone gives 0.005; decoding the true future graph (F) reaches 0.999. The tabular reference is higher at 0.605, or 0.46 when refitted without side count.

By grain class. Much of the pooled 0.411 is the sign of change that the side count fixes: by side count on validation the forecast reads 0.427 for fewer than six sides but 0.072 for six and 0.095 for more, and within the fewer-than-six and more-than-six classes it stays below the covariate reference (0.499 and 0.371).

Table S33: Grain validation by stage on 24,167 retained (pair, grain) cells from 756 pairs. Pearson R_{chg} and Spearman ρ pool recognised percentage-change cells. Brackets give 95% pair-bootstrap CIs; MAE is mean $\pm$ sample SD across cells, in percentage points.

Stage and arm	R_{chg} [95% CI]↑	ρ ↑	MAE (pp)↓
Stage 1 round trip, true future graph decoded (F)	0.999 [0.999, 1.000]	0.999	0.32±0.31
Stage 2, decoded request: **Deployed**, λ_c=1	0.447 [0.427, 0.466]	0.354	9.30±7.72
Stage 3, rendered: **Deployed**, p_{oracle}=0	0.411 [0.386, 0.435]	0.344	9.41±8.00

Measurement in the simulated domain. With exact simulated labels, the read-back reference reaches 0.993, against clinical readings of approximately 0.61 to 0.90 (Table S31 and Appendix C.2); the test failure above is the remaining read-back error on generated images.

In the selected qualitative example (Figure S5c), GraphMorph's rendered median grain-area change is within one percentage point of its request in each interval, against measured changes of $+4.2$ to $+8.7\%$; in its first row the labeller finds 536 grains for Pix2Pix, compared with 160 in the truth and GraphMorph.

G Representation fidelity and graph diagnostics

The anatomical representation returns the measured between-visit change at $R^{(32)}_{\mathrm{chg}} = 0.966$, and the forecaster predicts it at 0.366 over 32 regions and 0.448 on the core; both are read on 689 validation pairs, and test analyses are labelled.

G.1 Preservation of longitudinal change

The Stage 1 round trip preserves between-visit change. The deployed encoder returns the measured change over the 32 regions at $R^{(32)}_{\mathrm{chg}} = 0.966$ with slope 0.99 (Table S34). Pair by pair over the 32 regions, the change to be recovered is 2.81% of baseline volume and the round trip misplaces 0.74%, about a quarter of it, whereas the rendered and re-segmented forecast misplaces 2.62% on the same validation pairs and 2.67% on test (Table S23). The left hippocampus's change, about 32 voxels against a 12 mm latent site, returns at $R_{\mathrm{chg}} = 0.952$ because each site carries 16 channels that the decoder expands into a boundary; a one-bit occupancy grid at the same spacing returns it at $R_{\mathrm{chg}} = 0.26$ over the 689 validation pairs, and of the three structures probed the change survives only on the lateral ventricle (0.67 left, 0.68 right).

Change and surface distance are read alongside static Dice (Puglisi et al., 2025; Chen et al., 2026): both visits pass through the frozen autoencoder and are compared with the same 1.5 mm labels, and Dice is 0.947 for the 220-voxel accumbens and 0.753 for the 88k-voxel cortex. The primary pooled frame comprises seven structures in each hemisphere (14 regions): hippocampus, amygdala, lateral ventricle, thalamus, caudate, putamen and pallidum; 32-region and core results are labelled separately.

Spatial grid and latent resolution. Native 1.5 mm graphs return the measured change better than a graph built from 1 mm labels and resampled (the two-grid arm), and the deployed n_{ds}=3 grid is best on every metric of Table S34. The two-grid arm's reference is the label resampling itself: resampling labels segmented at 1 mm onto the 1.5 mm labels of the same scans, with no model involved, gives a fourteen-structure Dice of 0.9236 and a mean surface distance of 0.225 mm on 200 validation scans (0.908 over the 32 regions). On static shape the two-grid arm has little left to gain (Dice 0.9268, 0.218 mm), but on change the ordering reverses: it misplaces 1.77% of baseline volume against a 2.90% signal over the fourteen structures, against 0.71% for the native n_{ds}=4 arm with a third of the latent sites, and it is worse on 688 of the 689 pairs. Removing one further downsampling stage, the deployed native n_{ds}=3 arm reaches a fourteen-structure Dice of 0.9670 and a misplacement of 0.62%; cortical Dice rises from 0.635 to 0.753 and cortical surface distance falls from 1.021 mm to 0.464 mm. Only the native comparison, one lattice at two latent sizes, isolates latent capacity; the two-grid arm's learning-rate departure is recorded in Appendix I.7.

Table S34: Brain representation fidelity on 689 validation pairs from 186 subjects. Dice covers 32 regions and ASSD 16 structures; the resampled arm uses 60 pairs for both. ± is SD across 32 regions. Parentheses give latent-grid dimensions and site counts; MAE is in percentage points.

Arm	Dice ↑	ASSD (mm)↓	Slope	$R^{(32)}_{\mathrm{chg}}$ ↑	MAE$^{(32)}$ ↓
1 mm graph, resampled (12×14×12; 32,256)	0.888 ± 0.064	0.247	0.82	0.801 ± 0.111	1.78 ± 1.09
1.5 mm native, n_{ds}=4 (8×10×8; 10,240)	0.889 ± 0.092	0.229	0.97	0.953 ± 0.033	0.83 ± 0.53
1.5 mm native, n_{ds}=3 (16×20×16; 81,920)	**0.933** ± 0.055	**0.121**	**0.99**	**0.966** ± 0.026	**0.74** ± 0.56

How small a change the representation resolves. Typical change exceeds 2σ of the reconstruction error in 28 of 32 regions, and the four regions below this scale, the two accumbens areas and two cortices, still have round-trip change correlations of 0.899–0.915. Over the fourteen structures the deployed configuration has median/minimum SNR $3.1/2.4$, against $1.2/0.9$ for the two-grid arm.

One autoencoder for all regions. The encoder shares weights across regions, and the frozen decoder can express closer fits than the shared encoder produces: optimising each region's latent directly against the decoder on twelve validation pairs raises mean Dice from 0.932 to 0.982 over 32 regions, improving every region of every pair; the accumbens areas reach 0.994 and 0.997 and cortical Dice rises from 0.753 to 0.855. Doubling the budget from 2,500 to 5,000 steps on eight regions of three pairs changes Dice by 0.0009 on average and 0.006 at most.

G.2 Latent trajectories and region-pair biases

Figure S7 compares observed and predicted latent trajectories and their decoded regional changes.

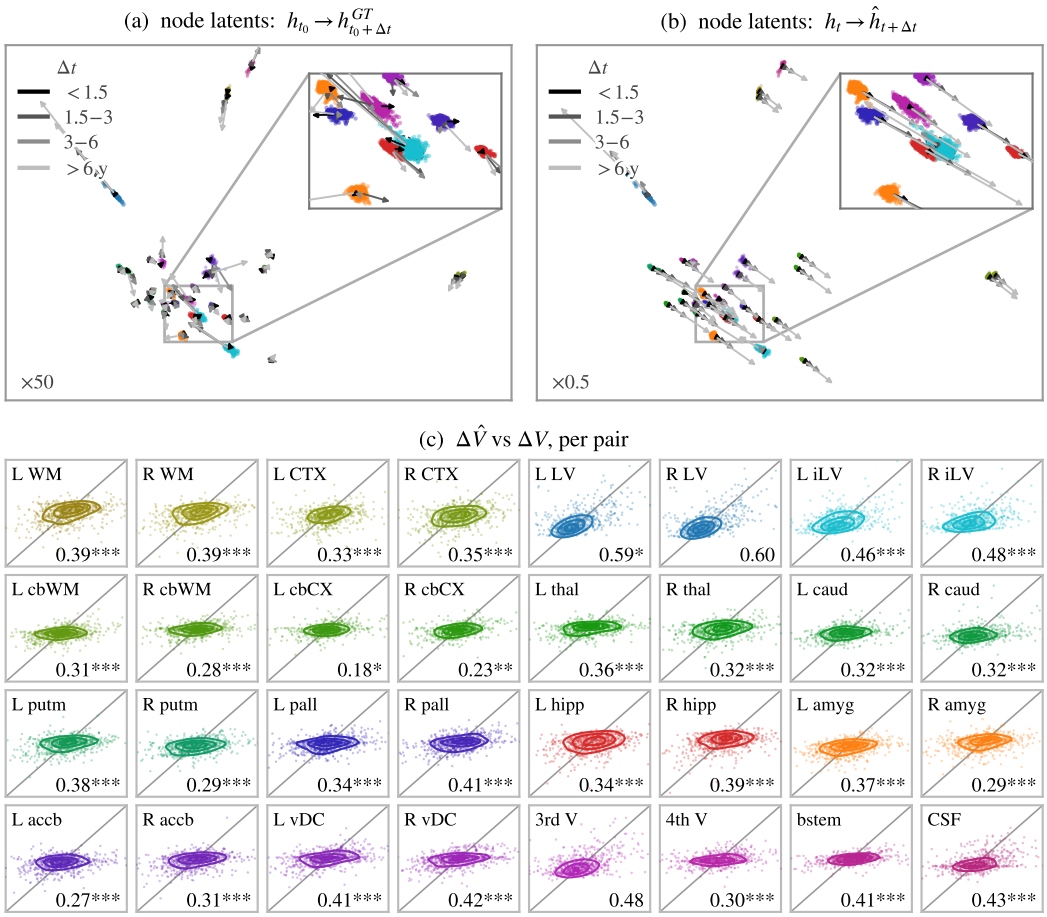

Figure S7: Latent trajectories and decoded anatomical change on validation. (a) Observed and (b) predicted latent trajectories after spatial pooling and two-dimensional PCA, with separate display multipliers; (c) decoded regional change.

The learned region-pair bias is an attention prior over anatomical regions, shared across subjects, ages and intervals; the full attention also depends on node features and conditioning. Adding a constant to any query row of B leaves attention unchanged, so the plotted signs describe the fitted bias parameters.

The prior's strongest block reproduces across reseeds and design variants. After scaling each run by its off-diagonal standard deviation without centring, 285 of 992 directed entries keep their sign across the six reseeds (median pairwise $r = 0.503$ over fifteen pairs) and 107 across all twelve runs (six reseeds, the deployed run and five variants; configurations in Appendix I.7). Figure 3 highlights the 44 pairs unanimous across the six reseeds with $|\bar{z}| \geq 1.6$: cortex, thalamus and basal-ganglia nodes have positive biases towards ventricular nodes, and inter-ventricular biases form the strongest negative block, consistent across all twelve runs. All four left-cortex-to-lateral-ventricle edges keep their sign across runs; the full cortex-to-ventricle block does not.

H EVALUATION PROTOCOL

H.1 EVALUATED IMPLEMENTATIONS

All compared methods are trained on this cohort. BrLP (Puglisi et al., 2025) and the MambaControl backbone (Yang et al., 2025) run in the latent space of the frozen medical-imaging autoencoder MAISI (Guo et al., 2025), the latter in the backbone configuration its release carries (labelled MambaControl in image panels). BrLP initialises ControlNet from its UNet. Δ-LFM (Chen et al., 2026) uses its author-provided autoencoder retrained on this cohort and trains its flow stage from scratch. DiffuseMorph (Kim et al., 2022) is the frozen backbone GraphMorph's renderer reuses, run with no control. CardiacAging (Campello et al., 2022) is a conditional adversarial age-progression synthesiser, and ImageFlowNet (Liu et al., 2025) runs in its neural-ODE form, with an ImageNet-pretrained ConvNeXt-tiny encoder used only for its latent loss. The four image-to-image baselines are Pix2Pix (Isola et al., 2017) and three further methods of Moschetto et al. (2025): Noise-Free Flow, Flow Matching and Diffusion. CardiacAging and these four baselines use no pretrained weights. Every prediction is scored as a reconstructed 3-D volume whatever the training dimensionality.

Tables 1 and 3 abbreviate model types as Deform (deformation), Noise (diffusion), ΔFlow (residual flow matching), Flow (conditional flow matching), ODE (rate integration) and Direct (image regression).

H.2 METRIC DEFINITIONS

Image fidelity. Within the same follow-up brain mask, prediction and ground truth are independently normalised to $[0, 1]$ using their respective 1st–99th percentiles. PSNR and SSIM use the whole volume, approximately three-quarters background. Masked-score sensitivity is reported in Appendix D.3.

Regional accuracy. Superscripts identify regions or aggregates: 32 denotes the brain graph regions; core the bilaterally merged hippocampus (Hipp), amygdala (Amyg) and lateral ventricle (LV); $\mathrm{core6}$ retains their six separate left/right regions. For the knee, 3 denotes femur, joint space (JS) and tibia. Brain masks are SynthSeg parcellations, with labels rounded to the nearest integer before thresholding (Appendix J).

Let p_{ir} and g_{ir} denote predicted and measured percentage changes for pair i and region r, using the stated baseline for each reading. Mean Pearson change correlation over brain or knee region set A is $R^{(A)}_{\mathrm{chg}} = |A|^{-1} \sum_{r \in A} \mathrm{corr}_i(p_{ir}, g_{ir})$. $\mathrm{MAE}^{(A)}$ averages the regions' mean absolute percentage-change errors, in percentage points (pp). Dice measures mask overlap. For reconstruction diagnostics, $|\Delta|$ is mean absolute measured change, σ is the SD of decoded-minus-measured change (both in pp), and SNR is their ratio. Bias is mean predicted-minus-measured change.

Regional progression errors. Progression errors use the measured baseline. Retain regions with $|g_{ir}| \geq \tau_i$, where $\tau_i = \frac{1}{32} \sum_r |g_{ir}|$, using the same threshold for the core. For $A \in \{32, \mathrm{core}\}$, $\langle \cdot \rangle_{i,A}$ averages retained structures; pairs with no retained structure are excluded. We measure relative mean absolute error (Δ_{RMAE}) and amplitude-matched symmetric regional error (Δ_{Region}):

$$\Delta^{(A)}_{\mathrm{RMAE},i} = \frac{\langle |p - g| \rangle_{i,A}}{\langle |g| \rangle_{i,A} + \epsilon}, \qquad \Delta^{(A)}_{\mathrm{Region},i} = \left\langle \frac{|g - sp|}{(|g| + |sp|)/2 + \epsilon} \right\rangle_{i,A}. \tag{6}$$

The single scale per method is $s = (\mathrm{mean}_i \langle |g| \rangle_{i,32} + \epsilon) / \mathrm{mean}_i \langle |p| \rangle_{i,32}$, estimated on its test predictions and shared with the core. Copy-forward uses $s = 1$; $\epsilon = 10^{-6}$. Cohort and clinical/history strata retain this full-test scale. Independent-segmentation comparisons estimate s within each segmenter's 100-pair subset.The test-fitted Δ_{Region} measures retrospective amplitude-matched error; raw MAE uses unscaled predictions.

Tables 3 and S27 use 7,517 strict-QC test pairs for joint-space correlation, raw MAE and image measurements, with stored masks defining measured change. Their joint-space progression errors use 2,850 read-back pairs, measured through the segmenter at both visits; three-region progression uses the full 7,517-pair frame. Three-region scoring thresholds at the within-pair mean absolute change and joint-space scoring at the frame mean absolute change, with a separate scale per method

and frame. The unrendered tabular reference predicts stored-mask change without conversion to the segmenter baseline.

Grain summaries use each method's identifiable grains. Their within-pair mean absolute measured change defines the threshold, and the scale uses the retained support. Read-back and exact-label measurements use their respective baselines. Paired comparisons recompute the threshold and scores on jointly identifiable grains with scales fixed.

Request adherence and calibration. Let V_k denote the volume (area for 2-D tasks) of region k, $\hat{S}$ the re-segmented render, Q^1 the decoded request and S^0 the measured baseline label map. $\mathrm{MAE}^{(A)}_{\mathrm{adh}}$ averages $100|V_k(\hat{S}) - V_k(Q^1)|/V_k(S^0)$ over (pair, region) cells over A; $R^{(A)}_{\mathrm{adh}}$ correlates rendered and requested change per region before averaging. Image-level request agreement also includes re-segmentation error. End-to-end change instead compares the render with the measured follow-up. For request accuracy, q^{decoded} measures change from the request's own decoded baseline and q^{common} from the measured baseline.

Agreement and calibration slopes fit $p = \alpha + \beta g$ and $g = a + bp$, respectively. For intercept-inclusive OLS on the same observations and weights, $\beta b = r^2$, where r is the individual-region correlation. Linear calibration requires $a = 0$ and $b = 1$; a low agreement slope alone does not establish miscalibration.

The identity reference is diagnosis-independent. Soft- and hard-mask request references are distinguished below. Request-response slopes regress realised on requested change and are distinct from β and b.

Boundary measurements. Surfaces are label-map voxels outside their 6-connected erosion at 1.5 mm isotropic spacing. Average symmetric surface distance (ASSD) uses the exact Euclidean distance transform with equal weight per surface voxel. Surface Dice $\mathrm{SDice}^{(A)}_{d}$ measures the fraction of surface voxels within d mm of the other surface, at $d = 1.5$ (one voxel, fixed before scoring) and secondarily $d = 3$. For rendered-image endpoints, both metrics compare the re-segmented render with measured follow-up labels. Means pool (pair, region) cells; empty predicted masks are recorded as failures and excluded from the means.

Knee image measurements. PSNR is measured inside the true follow-up bone region, SSIM across the image, and bone Dice averages femur and tibia agreement with the stored mask. SSIM and bone Dice are reported as percentages. Joint-space width is measured at one quarter and three quarters of the tibial plateau; JSW MAE scores the narrower compartment, in pixels of the 256×256 grid. No pixel spacing is recorded. Qualitative radiographs use anatomy-based intensity windows. The registration diagnostic restores the follow-up image to DiffuseMorph, while read-back measures the true follow-up through the evaluation segmenter against the stored masks. Request ASSD targets the stored follow-up mask. Render ASSD targets the segmented follow-up image.

Grain measurements. Grain readout is specified in Appendix J.3. R_{chg} is the Pearson correlation between read and true percentage change pooled over (pair, grain) cells; ρ is its Spearman counterpart. Constant forecasts have undefined correlation. Grain test uses 1,548 pairs with method-specific recognised cells. Validation baseline-mask Stage 1 reconstruction uses 24,192 mask–grain cells from 756 baseline masks; Stage 2 and Stage 3 use 24,167 retained pair–grain cells from 756 pairs.

Tabular references. Tabular predictors use elapsed interval, age, sex, diagnosis, history availability, baseline log volumes and previous-interval rates, with volumes counted from the 1.5 mm labels. They are fitted on training pairs. All-region tabular (Table S23) uses all 32 regions' volume and rate features, with ridge or gradient boosting selected per target on validation. Core-feature tabular (Table 1) uses the three core structures' features with gradient boosting alone. Table S8 uses the all-region configuration for both aggregates. So does the tabular row in Table S2. The merged-region tabular reference of Appendix A.2 selects ridge or boosting per structure using volume and rate features of seven bilaterally merged structures and CSF. Validation also selects the interval-only estimator. The tables' *bounds* are empirical references. For the elapsed-interval reference in Table 1, each core target is fitted to elapsed time on test. The validation-selected reference applies the

interval-only estimator chosen on validation to the test pairs (Table S8); it is the stronger of the two, and stars and paired contrasts use the test fit.

Renderer ablations. Table 4 sets the control map to zero, substitutes an identity evolver to request no change, or retrains with a projected latent increment or standard denoising loss plus a scalar control term. The full model uses decoded occupancy and the dense alignment objective.

Mask-route controls. The interval-plus-rate reference uses the interval, history indicator and each region's previous rate. For the brain mask route, predicted percentage changes are applied to measured baseline region counts, with targets rounded to the nearest nonnegative integer. A signed-distance adapter adjusts boundaries within three voxels, resolves overlaps with background as reserve, and passes the resulting masks through the same frozen renderer. The same-content graph control preserves predicted percentage changes, not absolute decoded volumes; one of its pairs exceeded a target count by one voxel and was retained. The deployed brain and knee renderers were trained on graph requests. For the knee, articular-margin shifts are split equally between femur and tibia to match the predicted joint-space area; the remaining bone boundaries are then adjusted to their area targets, with smoothed tie-breaking. The masks pass through the Stage-1 round trip before rendering.

Population and history requests. Population-trend requests use diagnosis-specific regional annual rates fitted through the origin on each diagnosis group's training pairs with equal subject weights: $\mu_{gr} = \sum_i w_i \Delta a_i g_{ir} / \sum_i w_i \Delta a_i^2$. The request is $\mu_{g(i)r} \Delta a_i$, using baseline-visit diagnosis. Direct-history requests extrapolate the previous percentage rate, clipped to $\pm 25\%$ per year; no-history pairs receive zero. The primary history comparison uses the same 503 validation pairs from 155 subjects, with 186 first pairs as a separate stratum. The history endpoint was fixed after the plan preview; the same-mask-route population decomposition is post hoc.

Registration controls. The registration control uses the SyN transformation (Avants et al., 2008) implemented in ANTsPy 0.6.3 (Tustison et al., 2021). It registers 32 independently thresholded (> 0.5) channels of the decoded baseline to those of the visit-coded forecast on the $120 \times 144 \times 120$ grid, using mean-square label distances, SyN $(0.2, 3, 0)$ and iterations $(100, 70, 50, 20)$ at shrink factors $(8, 4, 2, 1)$. The fixed transform warps the baseline image with linear interpolation. The learned renderer instead uses soft occupancies on its training canvas. The reconstruction anchor registers measured baseline labels to the decoded baseline; it is not a forecast, and its contrast with forecast registration changes both registration endpoints. Each arm's adherence is measured against its own request.

Registration realised a median 71% of requested volume change (within-pair ratios over regions with a hard-mask change of at least ten voxels, summarised across the 684 eligible pairs; the other five pairs remain in all 689-pair image endpoints). This is label-level realisation, separate from image-level adherence to soft volumes (summed occupancy) or hard volumes (thresholded voxel counts). No folding was observed under the fixed SyN settings.

Supervision-response controls. Supervision-response analyses use scales $\{0, 1, 2.5\}$ on brain and $\{0, 0.5, 1, 1.5, 2, 2.3\}$ on knee. Pooled response slopes include an intercept; within-pair slopes remove pair offsets. Directional monotonicity requires all successive responses to follow the reference direction. The identity arm uses the decoded baseline. Reference-request agreement and error against measured follow-up are reported separately.

H.3 Reporting conventions

Control and progression-error contrasts use 10,000 paired cluster-bootstrap draws (seed 0), resampling brain subjects, knee patients with both knees and all visits, or grain samples. ASSD levels in Table 5 and their paired contrasts share the subject draws. Pair-bootstrap and knee-bootstrap resample pairs and knees, respectively. Matched-seed intervals condition on the evaluated seeds. Progression-error differences use percentile 95% intervals with scales fixed. Scale-sensitivity intervals resample the full test cluster frame and refit test scales; validation scales are refitted on independently resampled validation clusters (seed 1). Retained progression support need not contain

every resampled cluster. Between-seed SD uses $ddof = 1$ and is reported separately. In progression blocks, † marks a difference from GraphMorph whose interval excludes zero, and each paired table specifies the subtraction order. An interval containing zero is not evidence of equivalence.

For brain R_{chg}, R_{adh}, volume and request-agreement MAE, and Stage-1 Dice/SNR, $\pm$ denotes SD across regions or core structures (denominator their number); brain PSNR, SSIM and progression errors use sample SD across pairs, and progression errors are pair means whose SD uses the $n - 1$ denominator. In Table S31, correlation $\pm$ values are pair-bootstrap SDs and MAE $\pm$ values are per-pair SDs; test brackets are 95% pair-bootstrap intervals, and oracle-control and read-back rows retain mean $\pm$ bootstrap SD. The spatial and supervision controls use patient-bootstrap intervals. Seed summaries, including the means in Tables S13 and S14, use between-seed SDs. Correlations and progression errors use three decimals; MAE, PSNR and SSIM use two. Bold marks the best unrounded mean in each model or training-configuration comparison; reference, (F), reseed and input-substitution readings are unranked, and Table S18 also includes copy-forward.

The elapsed-interval reference of an aggregate is the mean of its members' references, each the absolute correlation across pairs between the elapsed interval and that member's measured change. Brain benchmark stars test a method's R_{chg} against the elapsed-interval reference of the same aggregate by a one-sided paired subject-bootstrap over the 185 test subjects, each drawn with all its pairs, 10,000 resamples with seed 0 ($^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$, uncorrected); Table 4 tests full-model-minus-variant R_{chg} differences the same way. Table S22 reports one-year mean core change with $\pm$ half a 95% Wald interval. Stars mark realised-minus-occurred differences whose intervals exclude zero; daggers mark subject-bootstrap correlation differences from Δ-LFM whose intervals exclude zero. Table S26 also reports pair-bootstrap intervals and one-sided p-values; panel (b) reports subject-interval midpoints. The star assignments are unchanged between the two resampling units. Each core structure and the temporal horn are tested the same way against their own merged-pair reference (Appendix D.2); a mean can clear the mean of the references while one of its members does not.

Spearman's ρ agrees with Pearson's R_{chg} within 0.045 on the 32-region mean and 0.064 on the core for every main-table method. Bonferroni correction across the 22 tested cells (11 methods, two aggregates) retains only GraphMorph on both aggregates. Four compared readings exceed an individual structure's interval reference before correction, all on the amygdala ($p = 0.01$ to 0.026); none exceeds the core reference.

Domain-specific spreads and intervals. Boundary spreads are sample SDs across regional means. Knee $R^{(3)}_{\mathrm{chg}}$, $\rho^{(3)}$ and $\mathrm{MAE}^{(3)}$ use regional SD; individual-region correlations use pair-bootstrap SD (10,000 draws, seed 42; denominator 10,000) with one-sided stars against the elapsed-interval reference; knee regional MAE uses sample SD of per-pair absolute errors, and PSNR, SSIM and reconstruction checks sample SD across pairs or masks. Grain area MAE uses sample SD over (pair, grain) or (mask, grain) cells, and grain correlation SDs and intervals a pair bootstrap (10,000 draws, seed 42). Tables S10 and S11 use paired subject-bootstrap intervals. Cohort summaries report age as mean $\pm$ SD, sizes and intervals as medians (IQR or range), and counts as n (%) of the stated unit.

Independent segmentation. The independent-segmenter subset was fixed before reading model outputs (seed 0; ADNI 80, OASIS-3 15, AIBL 5; one pair per subject). A prespecified mapping folds FastSurfer v2.5.4 DKT labels (Henschel et al., 2020) into the 32 regions: cortical parcels into cortex, choroid plexus into lateral ventricles, and white-matter hypointensities into the nearer hemisphere's white matter. SynthSeg's extracerebral CSF and FastSurfer's ventricular-adjacent CSF differ anatomically and are not directly comparable. Independent-segmentation intervals use 10,000 paired subject-bootstrap draws.

The compatible 29-region map (compat29) excludes CSF and merges cortex with white matter in each hemisphere. The noCSF31 map excludes CSF alone. The compat27_vent map also merges the lateral and inferior lateral ventricles within each hemisphere. The 32-region map keeps the full SynthSeg label set. The core3 map contains the three core structures. Pooled samples combine the 100-pair subset and the independent sample (396 pairs, 185 subjects). The independent sample contains 85 subjects outside the 100-subject subset. Four Δ-LFM segmentations fail, leaving 292

complete pairs from 84 subjects; the failure-as-copy-forward sensitivity retains 296 pairs. Consensus change is the mean of the two segmenters' percentage changes.

For same-image volumes X (SynthSeg) and Y (FastSurfer), concordance is $\mathrm{CCC} = 2\,\mathrm{cov}(X,Y)/[\mathrm{var}(X)+\mathrm{var}(Y)+(\bar{X}-\bar{Y})^2]$. Volume bias is the mean of $200(X-Y)/(X+Y)$, in percent. Map-level change-agreement rows average the per-region correlations and concordances; they are not pooled correlations or measurement ceilings.

Measurement-ceiling estimator. The regional ceiling is $\sqrt{\max(0,\, 1-2s^2_{\mathrm{seg}}/\mathrm{var}\,\Delta\log V)}$. Here $\Delta\log V$ is consecutive-visit change and s_{seg} is the pooled residual of per-subject linear fits of $\log V$ on age. The factor two accounts for the two segmentations in each change. Test fits use 154 subjects with at least three visits (792 scans), with change variance estimated over all 669 pairs from 185 subjects. The validation estimate (Appendix A.2) uses 155 subjects (813 scans) and all 689 pairs from 186 subjects.

H.4 Calibration and subgroup analysis specifications

Validation calibration uses 689 pairs from 186 subjects: CN 233/60, MCI 319/71 and AD 137/55 (pairs/subjects), with 95% subject-bootstrap intervals. Prediction-bin boundaries are the frozen quintiles of the earlier validation forecasts, applied unchanged with open outer bins, so the bins need not be equally populated.

The magnitude analysis uses 669 test pairs from 185 subjects and fits the mean $100|\Delta V|/V_0$ over the three core structures on Δa and Δa^2, with subject-level random intercepts and slopes, by restricted maximum likelihood (fits selected by likelihood across four optimisers). The linear rate analysis drops the quadratic term and compares covariate-adjusted and within-/between-subject specifications (Appendix C.5); the sex-only adjustment uses 568 pairs, and the all-three adjustment retains 669 pairs by coding missing sex as a third level. Rate-difference intervals use 400 paired subject-bootstrap resamples (seed 11), refitting the REML models with L-BFGS. Pair-level image-fidelity associations are also adjusted for interval.

For signed change at one year, measured and rendered outcomes are stacked and fitted as $y \sim (\Delta a + \Delta a^2) \times \mathrm{reading}$, with subject-level random intercepts and slopes and a pair variance component; the lateral-ventricle specification was fixed on validation before the test read and the core outcome added afterwards. Stratum contrasts use the pooled interaction fit rather than subtraction of printed levels; companion results cover hippocampus, amygdala, the interval-band and ADNI-only sensitivities on both splits, and Δ-LFM uses the same model.

H.5 The follow-up channel and F-ablation

On brain test, DiffuseMorph reaches core correlation 0.540 with the follow-up scan in the score network's conditioning channel and 0.077 with that channel withheld from the same backbone (Tables S24 and 1). Request substitutions are defined in Appendix I.5.

In the released DiffuseMorph baseline inference path (Kim et al., 2022), the score network receives $[M, F, F]$ and the deformation network receives $[M, \gamma \cdot \mathrm{score}]$ at a fixed timestep without sampled noise. Our F-ablation uses the same checkpoint and test pairs with only the follow-up channel replaced. Restoring that channel changes PSNR by more than 0.5 dB on 495 of 669 pairs, with a mean increase of $+1.05$ dB.

H.6 Knee inputs, sampling and model selection

What each method receives. On the knee, history reaches only GraphMorph and the tabular reference. ImageFlowNet uses its local spatial-graph variant ImageFlowNet-GraphODE for knee and grain data and its ODE variant on the brain. GraphMorph receives the baseline radiograph rescaled to $[-1, 1]$, its stored mask, the previous visit's mask where one exists, the elapsed interval, the baseline age, the knee's Kellgren–Lawrence grade collapsed to three levels (0–1, 2, 3–4) as a one-hot vector, and the preceding pair's per-region rate of area change (clipped to $\pm 25\%$ per year) with an indicator that one exists. Sex never reaches the forecast, the follow-up grade is replaced by the visit-00 grade before the loss and is not passed at inference, the score network's follow-up slot holds the

baseline image, and the follow-up mask enters only the substitution arm. ImageFlowNet-GraphODE receives the baseline radiograph, the interval in months, the same three-level grade and the baseline age, with no history image; DiffuseMorph receives the baseline radiograph alone under the same flag, and the four image-to-image baselines the baseline radiograph alone without the elapsed interval (Table S37(b)). The grade is the knee's OAI visit-00 reading, used by GraphMorph, the tabular reference and ImageFlowNet-GraphODE; for 28 ungraded knees, GraphMorph and the tabular reference use KL 0 and ImageFlowNet-GraphODE its highest severity category, and excluding the 15 affected test pairs changes each $R^{(\mathrm{JS})}_{\mathrm{chg}}$ by less than 0.001.

Sampling. Every row is one prediction per pair with no sample averaging. GraphMorph, DiffuseMorph, ImageFlowNet-GraphODE and Pix2Pix are deterministic (the score network is evaluated once at $t{=}0$ and the deformation read after a seven-step interpolation of its output; the ODE uses fixed-step Runge–Kutta). Noise-Free Flow integrates its velocity field with 30 Euler steps from the baseline image without noise; Flow Matching takes 30 Euler steps from an unseeded Gaussian draw concatenated with the baseline, and Diffusion runs its scheduler's full $1{,}000$ DDPM steps from an unseeded Gaussian draw conditioned on the baseline, its output percentile-normalised before it is written.

Training budgets and checkpoints. Each knee benchmark row uses one training run on one GPU under its stated budget. GraphMorph uses the 25,691 registration-gated training pairs and deploys the last checkpoint after 60, 24 and 20 epochs for its three stages; the other methods use ungated data. Pix2Pix trains for 40 epochs at batch 32 on all 26,196 pairs, selecting epoch 13 by validation L_1; both flow baselines train for 20 epochs at batch 4 on the same pairs, with validation picks on all 6,731 validation pairs at epoch 20 (Noise-Free Flow) and 19 (Flow Matching); Diffusion trains for 20 epochs on 2,619 training pairs, then 10 and 30 epochs on the full set at batch 8, deploying epoch 1 of the final phase. Pair presentations are 523,920 for each flow baseline, 1,047,840 for Pix2Pix, about 1,100,000 for Diffusion and 513,820 for GraphMorph's Stage 3. ImageFlowNet-GraphODE trained for 11 of 15 epochs with 2,048 sampled ordered pairs per epoch (22,528 presentations), including non-consecutive visits, each draw selecting a training knee and then a pair of its visits; validation prediction PSNR on up to 256 items per epoch selected epoch 8 with patience three. This variant uses its spatial graph without the change loss or history image.

Knee tabular reference. A gradient-boosted regressor is fitted separately per region on QC-passing training pairs. Inputs are elapsed interval, first-visit age, sex, body-mass index, side, Kellgren–Lawrence grade, a history indicator, nine baseline geometry measurements (three regional areas, three joint-space width statistics and three compartment areas), and their previous-interval rates. Missing inputs use training medians. Grade uses the visit-00 reading described above; its removal and the other feature sensitivities are reported in Appendix E.

H.7 Evaluation scope and selection

Prespecified test evaluations cover the no-correlation-loss and no-history-input models (seed 44) and the independent FastSurfer sample (compatible 29-region primary, complete cases). Compatible-ROI rescoring of the 100-pair subset and cross-measurement analyses are post hoc. Clinical-stratum comparators are selected by test $R^{(32)}_{\mathrm{chg}}$, whereas the AD validation comparison selects its comparator on validation.

The deployed brain seed 42 was fixed beforehand and the deployed knee setting was selected on validation before rendering. Brain renderer ablations and reseeds 43–45 are evaluated on test, with the reseeds on the training canvas. The choices that touched test data are reported with their sensitivity: the knee Dice 0.85 QC threshold was selected on test and Appendix E reports filtered and unfiltered results, and the grid, oracle-schedule and control-signal probes used test data before their validation analyses. Grain comparisons use fixed budgets without model selection.

Figure 2 and its complete-method counterpart use core-error separation among subjects with at least three test pairs. The random panel uses seed 0 among subjects with at least three consecutive pairs. Progression and control examples use agreement with observed trajectories. The forecasting-failure example maximises subject-mean core request error. The rendering-failure example maximises render/request discrepancy after excluding that subject. The knee example is a multi-visit test knee,

the architecture illustration uses one test pair, and the grain example is the validation microstructure with the largest deployed versus no-control-map separation.

Qualitative display. In Figure 2, brain-mask intensity percentiles 1 and 99 are computed separately for each image; left-panel absolute errors share a zero-to-99th-percentile scale pooled over all displayed methods and visits (copy-forward included), and right-panel change $\hat{F}_t - \hat{F}_0$ is relative to the zero-interval render from the same age-79 input, with its own symmetric pooled 99th-percentile scale. The complete-method and random-subject panels each use a shared residual scale.

I METHOD AND REPRODUCIBILITY DETAILS

Notation follows Section 3. All GraphMorph stages use Adam with cosine learning-rate decay to 10^{-6}; epochs, batch sizes and initial learning rates are in Table S35.

Table S35: GraphMorph training settings: epochs / batch size / initial learning rate. Each stage uses its final epoch; architecture and losses are specified in Appendix I.

Stage	Brain	Knee	Grain
1: anatomical autoencoder	$100 / 8 / 6{\times}10^{-4}$	$60 / 16 / 6{\times}10^{-4}$	$60 / 16 / 6{\times}10^{-4}$
2: evolver	$24 / 2 / 2{\times}10^{-4}$	$24 / 16 / 2{\times}10^{-4}$	$24 / 16 / 2{\times}10^{-4}$
3: renderer	$20 / 1 / 2{\times}10^{-4}$	$20 / 8 / 2{\times}10^{-4}$	$20 / 8 / 2{\times}10^{-4}$

I.1 STAGE 1: ANATOMICAL AUTOENCODER

Given a segmentation S with $K{=}32$ labels, the encoder E maps each region's binary mask to a spatial latent,

$$Z = E(S) \in \mathbb{R}^{K\times c\times d\times h\times w}, \tag{7}$$

with $c{=}16$ channels, eight base encoder channels, and three stride-2 convolutions, so the padded $128{\times}160{\times}128$ canvas (Appendix I.7) gives a $16{\times}20{\times}16$ grid per region. Group normalisation removes absolute scale, so a size embedding computed from the mask occupancy, the *volume token*, is added back as a channel bias. The graph encoder G mixes the K nodes through two graph-attention layers (four heads, node width 64) with FiLM conditioning and returns $H = G(Z)$ of the same shape by a residual update. The decoder D emits per-region occupancy logits at the input resolution through three transposed convolutions. The stage trains as an autoencoder on single segmentations,

$$\mathcal{L}_1 = \mathcal{L}_{\text{seg}}\big(D(G(E(S))), \mathbb{1}[S{=}k]\big) + \lambda_{\text{vol}}\big|\log(\hat{V}_k + 1) - \log(V_k + 1)\big|, \tag{8}$$

where $\mathcal{L}_{\text{seg}}$ is soft-Dice plus binary cross-entropy, $\hat{V}_k$ is the decoded occupancy mass of region k and V_k the true one, and $\lambda_{\text{vol}}{=}1$.

Each brain step decodes eight sampled regions. The knee and grain encoders retain $c = 16$ channels and use four stride-2 levels, giving a 16×16 latent grid; each step decodes all three knee or all 32 grain regions.

I.2 STAGE 2: FORECASTER

With E, G and D frozen, the evolver $\mathcal{P}$ advances the graph over the interval, $\hat{H}_F = \mathcal{P}(H_M, \Delta a, b, q)$ with $H_M = G(E(S_M))$. A node is summarised by the spatial mean of its latent, 16 numbers, concatenated with its region identity and embedded to 128 dimensions. Two blocks of graph attention, four heads each, mix the 32 summaries. FiLM conditioning precedes each block, computed from a sinusoidal embedding of the integration step Δt, baseline-visit diagnosis, and the subject's age advanced at each sub-step. When a prior visit exists, two history features q join the summaries before the first block: the pooled latent velocity between that visit and the baseline, and that interval's per-region rate of change. Pairs without a prior visit carry neither. A learned

32×32 bias, shared across the heads, is added to the attention logits. A linear read-out of the mixed summary gives γ_k and β_k, 16 of each, and a zero-initialised convolution emits the increment. The increment is integrated in $n = \min\{12, \max[1, \lceil \max_{i \in \mathcal{B}} |\Delta a_i| \rceil]\}$ Euler steps shared by minibatch $\mathcal{B}$, with intervals in years and $\Delta t_i = \Delta a_i / n$. A zero interval leaves the latent state unchanged. Brain decoder-space evaluation and end-to-end inference process one pair at a time. The deployed evolver trains in batches of two. All 32 nodes carry the update, and cross-region information reaches a node's latent only through those 32 numbers per step. The heads are concatenated and projected together and share the bias, so head identity plays no part in training or inference.

The objective combines latent supervision with decoded occupancy, volume and correlation losses:

$$\begin{aligned}\mathcal{L}_2 = \lambda_g \,\mathrm{SmoothL1}\big(\hat{H}_F, G(E(S_F))\big) + \lambda_s \,\mathcal{L}_{\mathrm{seg}}\big(D(\hat{H}_F), \mathbb{1}[S_F{=}k]\big) \\ + \lambda_v \,\big|\log(\hat{V}_k + 1) - \log(V_k^F + 1)\big| + \lambda_c \,\big(1 - r_k\big),\end{aligned} \tag{9}$$

with the decoder-supervised terms evaluated on a random eight regions per step (Appendix A.2) and λ_g=1, λ_s=10, λ_v=20, λ_c=1. For knee, decoder supervision covers all three regions and $\lambda_c = 0$; grain samples three of its 32 regions per step with $\lambda_c = 1$.

For each decoded region k, r_k correlates predicted and measured relative volume changes read through the frozen decoder over the current and previous 16 batches. Gradients pass through the current batch only.

I.3 Stage 3: renderer

For the sampled regions $\mathcal{K}$, let $w_k = \mathcal{T}(\mathbb{1}[S_M = k], \phi)$ be the warped baseline mask and $o_k = O_{\Delta a}[k]$ the requested occupancy. The renderer's anatomical loss $\mathcal{L}^{\mathrm{anat}}_{\mathrm{render}}$ of Eq. 5 is the alignment objective

$$\mathcal{L}_{\mathrm{align}} = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \left[1 - \frac{2\langle w_k, o_k \rangle + 1}{\|w_k\|_1 + \|o_k\|_1 + 1} + \lambda_{av} \big| \log(\|w_k\|_1 + 1) - \log(\|o_k\|_1 + 1) \big| \right], \tag{10}$$

Each brain step samples eight regions at random, with λ_{av}=10. Knee aligns all three regions and grain samples three of 32 per step. The target o_k is the frozen anatomical prediction. The total objective is

$$\mathcal{L}_3 = \lambda_a \mathcal{L}_{\mathrm{align}} + \lambda_L \lambda_{\mathrm{sim}} \,\mathcal{L}_{\mathrm{NCC}}(M \circ \phi, F) + \lambda_L \,\mathcal{L}_{\mathrm{smooth}}(\phi), \tag{11}$$

with λ_a=1, λ_L=20 and fixed λ_{sim}=0.005.

Renderer configuration. Stage 3 trains only the control projection C and control branch. The control field has 32 channels. The brain score network has base width 8 and multipliers $(1, 2, 4, 4)$, with field-decoder widths $(32, 32, 32, 8, 8, 3)$. Knee and grain use base width 16, multipliers $(1, 2, 4, 8, 8)$ and field-decoder widths $(32, 32, 32, 16, 16, 2)$. Brain Stage 3 uses no gradient-norm clipping; knee and grain clip at 1.0. Inference follows Section 3.2. Parameter counts and canvas definitions are in Appendix I.7.

Architecture illustration. Figure 1 uses deployed intermediates from one test pair spanning 7.9 years: requested left-lateral-ventricle change is $+33.5\%$, against measured $+38.8\%$. Anatomical 2-D panels share the axial slice at the 65th percentile of labelled voxels along z; the latent mosaic uses each region's middle slice. The attention probe averages validation subjects. The training band uses a separate pair during registration pretraining: q noises the follow-up, η scales deformation, $\mathcal{T}$ warps the image and NCC is local normalised cross-correlation.

I.4 Registration pretraining and the frozen renderer

The deformation wrapper is a three-dimensional DiffuseMorph (Kim et al., 2022) trained at 1.5 mm from random initialisation on the 5,331 consecutive training pairs, 14 epochs at $4{\times}10^{-4}$ followed

by 51 at 2×10^{-4}; GraphMorph freezes epoch 65. It is a registration model, so both scans of a pair are present at that stage: the follow-up is the diffusion target, a conditioning channel of the score network and the target of the image similarity term. The objective is the denoising loss plus, at weight 20 each, a local normalised cross-correlation between the warped baseline and the follow-up and a smoothness penalty on the displacement field. One optimiser covers both networks, so the registration terms also shape ϵ_θ.

Stage 3 reuses the frozen wrapper weights and uses the follow-up image for its photometric training term, weighted by $\lambda_L\lambda_{\text{sim}}$ rather than the pretraining weight 20. Inference replaces the follow-up image with the baseline. Restoring it yields the registration diagnostic in Table S24.

The knee wrapper. The deformation wrapper the knee Stage 3 freezes, shared with both DiffuseMorph rows, is a two-dimensional DiffuseMorph trained from random initialisation for 50 epochs at a fixed rate of 10^{-4} under the brain wrapper's objective on all $26{,}196$ consecutive training pairs (no registration gate, so it saw the 505 pairs the three GraphMorph stages exclude); the frozen checkpoint is the last. The wrapper holds $4{,}314{,}019$ parameters, none updated by Stage 3, which trains the control-map builder and the control branch ($1{,}614{,}880$ parameters); with Stage 1 at $195{,}313$ and the Stage 2 evolver at $473{,}209$, the three stages hold $6{,}597{,}421$.

I.5 Request substitutions

The request can be changed while the baseline and the renderer stay fixed. **Oracle control (F)** replaces $\hat{H}_F$ with $G(E(S_F))$, supplying measured follow-up anatomy through the same decoder; this follow-up-informed diagnostic retains segmentation and reconstruction error and changes the request distribution. **Identity evolver** sets $\hat{H}_F = H_M$, requesting no change from the decoded baseline anatomy. Because of reconstruction error the rendered image need not equal the baseline exactly. **Zero control** sets $U = 0$, removing baseline-anatomy and progression information alike from the control branch. **Scaled request** decodes $H_M + s(\hat{H}_F - H_M)$, with s=1 recovering the forecast. Because the decoder is nonlinear, scaling the latent increment need not scale regional volume change linearly. The sweep over s is in Appendix B.3.

I.6 Brain training budgets and deployment

Every brain method is evaluated on the same 669 consecutive test pairs; Table S36 lists training data, budgets and checkpoint rules. DiffuseMorph's epoch 65 was selected by registration gain on 40 validation pairs. Inference is one deterministic pass with no test-time free parameter for GraphMorph (training canvas $128\times160\times128$), DiffuseMorph (F-ablated, $F{:=}M$, scored at t=0), CardiacAging (conditioned on Δage, all 120 slices), ImageFlowNet (ODE to Δt, all slices) and Pix2Pix; BrLP samples 50 DDIM steps (one unseeded sample; follow-up covariates estimated by its disease-course model), the MambaControl backbone runs SDEdit from t=150 (16 executed denoising steps, seed 0, one sample), Δ-LFM 50 flow steps (seed 1234, two antithetic samples averaged), Noise-Free Flow 300 Euler steps from the baseline, Flow Matching 300 Euler steps from noise (one unseeded sample) and Diffusion 1,000 DDPM steps (one unseeded sample).

Table S36: Brain comparator training budgets and checkpoint selection. En denotes epoch n; inference and deployment settings are given in Appendix I.6.

Method	Training samples	Epochs × batch	Deployed checkpoint
DiffuseMorph	5,331 consecutive pairs	65 × 8	Validation-selected E65
BrLP	UNet: 6,806 scans; ControlNet: 15,666 ordered pairs; disease-course model on training data	UNet 20 × 8; ControlNet 100 × 8	Final epoch of each stage
MambaControl backbone	Diffusion: 6,806 scans; ControlNet: 15,666 ordered pairs	Diffusion 20 × 8; ControlNet 30 × 8	Final epoch of each stage
Δ-LFM	15,666 ordered pairs, up to two prior visits	9 × 8	Elementwise mean of E6–8 weights
CardiacAging	68,170 cross-sectional slices (6,817 scans × 10 slices)	100 × 64	Final E100
ImageFlowNet	2,048 random (subject, slice) draws per epoch; a random pair of any two visits	110 × 64	Best validation PSNR, E86
Pix2Pix	53,310 slice pairs (5,331 pairs × 10 slices)	30 × 32	Best validation L_1, E2
Noise-Free Flow	53,310 slice pairs	100 × 16	Validation-selected E64
Flow Matching	53,310 slice pairs	100 × 16	Validation-selected E75
Diffusion	53,310 slice pairs	100 × 16	Final E100

Δ-LFM's 4×4-voxel texture also occurs in its follow-up autoencoder round trip: the corresponding band contains 2.9% of high-frequency energy, versus 0.0% for MAISI at the same compression. Table S24 reports both reconstruction readouts.

MambaControl backbone. We use the released diffusion–ControlNet backbone without the Fourier anatomy-graph module. It predicts by SDEdit from t_{start} and blends the decoded change at strength α, $\hat{F} = \text{dec}(\text{enc}(M)) + \alpha\,(\tilde{F} - \text{dec}(\text{enc}(M)))$ with $\tilde{F}$ its decoded generation. The three test-time free parameters were frozen by a sweep over $t_{\text{start}} \in \{150, 200, 250\}$, $\{25, 50, 100\}$ DDIM steps and $\alpha \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$ on the first 60 of the 689 consecutive validation pairs (47 subjects), before test evaluation; the requested change correlates with the measured change at best 0.124, at the gentlest t_{start} and independently of α, and no grid point beats copy-forward on both PSNR and SSIM. The selected setting uses $t_{\text{start}} = 150$, a 100-step schedule and $\alpha = 0.10$. For the renderer-transfer experiment, the eight-dimensional context holds follow-up age, a constant, diagnosis and five regional volumes (cortex, hippocampus, amygdala, white matter and lateral ventricle), with baseline, decoded-forecast and measured-follow-up volumes mapped through the same affine maps ($R^2 = 0.970$–0.9997); all three contexts are evaluated on 689 validation pairs at $t = 150$, 100 steps and blend weights 0.1 and 1.

I.7 Training settings and reproducibility

Training budget and settings. Each deployed stage is one run on one GPU at a fixed seed (Table S35), and each stage freezes its predecessors; Stage 3 also freezes both DiffuseMorph networks. Brain Stage 1 and Stage 2 have $208{,}913$ and $484{,}400$ parameters; Stage 3 trains $278{,}688$ of the renderer's $1{,}355{,}652$, and the full pipeline has $2{,}048{,}965$.

Compute. GraphMorph stage training for the 1.5 mm brain, knee and grid experiments, including ablations and reseeds, totalled about 345 h. Runs used an NVIDIA RTX PRO 6000 Blackwell (96 GB), except the grain extension and the grid experiment's Stage 1, which used an RTX 6000 Ada (48 GB). The grid experiment's Stage 1 ran at batch 4 rather than 8 to fit its 56-region decoder, and its cell rule dropped seven of the $6{,}817$ training visits (none in validation or test).

Ablation training configurations. Three resumed arms differ from the deployed run in optimisation as well as in their named factor, and resuming restarts the sampler's random stream. Stage 2 comparisons fix the data and the Stage 1 checkpoint; epochs are numbered from 1. Full-32 decode uses batch 4 rather than 2. The resumed core-6 arm uses about 0.57 of the cosine learning rate at epochs 12–17 and as little as 0.14 thereafter through epoch 23. The original 32-region arm, distinct from the Baseline row of Table S2, resumes at epoch 10 and uses 0.63–0.83 of the prescribed rate at epochs 11–23. The Stage 1 two-grid arm resumes at epoch 20 and uses 0.90–0.98 at epochs 21–98.

History inputs and correlation loss. The Baseline row and the deployed model differ in history coverage, not in the correlation-loss buffer. History rates come from the previous and baseline segmentations; in the Stage 2 Baseline row all but one regional history-rate input were zero, whereas the deployed model and its reseeds use rates for all regions when a prior visit is available. For the correlation loss, the Baseline, deployed model and reseeds use region-matched training-history buffers (Appendix I); the original 32-region arm matched this buffer by column position. The wider-loss diagnostic uses four times the sample count and 64 history batches (Appendix A.2).

Seeds and the consensus pool. Stage 2 reseeds 44–49 appear in Table S2 and Stage 3 reseeds in Appendix B.4. The consensus pool (Appendix G.2) contains twelve Stage 2 runs: six reseeds, the deployed, Baseline, full-32 decode and no-Δt-scaling runs of Table S2, the original 32-region arm and a batch-4 replicate. Seed 45 resumes from epoch 12 with the schedule reset; its learning rates at epochs 13–24 match the five uninterrupted reseeds.

Inference canvas. Every reading is taken on the store grid $(120, 144, 120)$: inference zero-pads volumes to the training canvas $(128, 160, 128)$ and crops predictions back to the store grid. GraphMorph's test reads are rendered on this canvas. Validation readings marked as on the default canvas (Appendices A.4 and B.4) use padding to $(192, 224, 192)$; results from the two canvases are reported separately for each checkpoint.

Diagnosis input. The diagnosis supplied to the model is the one recorded at the pair's input visit, one-hot over CN, MCI and AD, in training, decoder-space evaluation and brain rendering alike. Identity requests are diagnosis-independent. Mask-route construction and the registration control are specified in Appendix H.2. Request-level seed means average metrics from seeds 44–46, not predicted images.

Knee-specific training. Stage 3 skips non-finite losses or gradients. At unit weight, the similarity term's gradient is about twice the alignment term's on the knee and a sixth on the brain; at weight 0.005, its knee contribution is about one per cent of the alignment gradient.

Grain-specific training. Grain uses the knee configuration with the task-specific settings in Table S35 and a DiffuseMorph backbone retrained on grain data. Each network is trained once at seed 42 (Table S33).

J Data and cohort construction

J.1 Brain cohort and grid

Cohort. Three studies pooled and split by subject give $1{,}857$ subjects (ADNI $1{,}315$; OASIS-3 356; AIBL 186) and $8{,}546$ scans (ADNI $6{,}752$; OASIS-3 $1{,}188$; AIBL 606); age, interval (Table S37(a)) and follow-up span match closely across splits, while source, sex and diagnosis shares differ by up to nine points, mostly in the smaller splits. Sex, age and per-visit diagnosis come from the upstream per-visit covariate table; each study's own state names (CN in ADNI, HC in AIBL, Normal in OASIS-3) are pooled to CN, MCI and AD, and no model receives the study of origin. Of the $1{,}500$ subjects with sex on record, 819 (54.6%) are female; sex is unrecorded for the 356 OASIS-3 subjects and one ADNI subject. By their last visit, 266 subjects (14.3%) had progressed to a more severe state. Intervals are right-skewed (test mean 1.03 years, longest 7.9), so the median of about seven months is quoted.

Table S37: Cohort characteristics by split. Age is mean ± SD; intervals, grain counts, side counts and area changes are medians with IQR or range as labelled. Counts use the stated row units; percentages refer to subjects in (a) and knees in (b). Two brain training subjects have no recorded diagnosis.

Characteristic	All	Train	Validation	Test
(a) Brain MRI				
Subjects, *n*	1,857	1,486	186	185
Consecutive pairs, *n*	6,689	5,331	689	669
Age at first visit, years	72.4 ± 8.0	72.3 ± 8.0	72.6 ± 8.1	72.4 ± 7.8
Inter-visit interval, years, median (IQR)	0.59 (0.44–1.08)	0.58 (0.44–1.09)	0.60 (0.47–1.07)	0.59 (0.44–1.06)
CN at first visit, *n* (%)	539 (29.0)	431 (29.0)	60 (32.3)	48 (25.9)
MCI at first visit, *n* (%)	736 (39.6)	591 (39.8)	71 (38.2)	74 (40.0)
AD at first visit, *n* (%)	580 (31.2)	462 (31.1)	55 (29.6)	63 (34.1)
(b) Knee radiographs				
Patients, *n*	4,206	2,680	685	841
Knees, *n*	8,295	5,286	1,344	1,665
Consecutive pairs, *n*	41,144	26,196	6,731	8,217
Age at first visit, years	61.0 ± 9.1	61.0 ± 9.1	60.7 ± 9.1	61.4 ± 9.3
Inter-visit interval, years, median (IQR)	1.00 (1.00–2.00)	1.00 (1.00–2.00)	1.00 (1.00–2.00)	1.00 (1.00–2.00)
KL 0 at first visit, *n* (%)	3,344 (40.3)	2,134 (40.4)	534 (39.7)	676 (40.6)
KL 1 at first visit, *n* (%)	1,496 (18.0)	921 (17.4)	260 (19.3)	315 (18.9)
KL 2 at first visit, *n* (%)	2,101 (25.3)	1,367 (25.9)	341 (25.4)	393 (23.6)
KL 3 at first visit, *n* (%)	1,107 (13.3)	699 (13.2)	177 (13.2)	231 (13.9)
KL 4 at first visit, *n* (%)	247 (3.0)	165 (3.1)	32 (2.4)	50 (3.0)
(c) Grain growth				
Microstructures (samples), *n*	1,200	816	126	258
Consecutive pairs, *n*	7,200	4,896	756	1,548
(Pair, grain) cells, *n*	230,400	156,672	24,192	49,536
Grains in the lattice at 0 MCS, median (range)	149 (144–150)	149 (145–150)	149 (144–150)	149 (146–150)
Grains in the lattice at 800 MCS, median (range)	80 (66–95)	80 (66–95)	81 (70–92)	80 (66–92)
Baseline side count of a tracked grain, median (IQR)	7 (6–8)	7 (6–8)	7 (6–8)	7 (6–8)
Per-grain area change over an interval, % (IQR)	6.1 (–2.3, 13.7)	6.1 (–2.3, 13.7)	6.1 (–2.3, 13.8)	6.1 (–2.2, 13.7)

Provenance of the grid. Images registered to MNI space at $121 \times 145 \times 121$ voxels and 1.5 mm isotropic spacing pass through an intermediate 1 mm lattice before resampling to the $120 \times 144 \times 120$,

1.5 mm generation grid; the band above the 1.5 mm Nyquist frequency holds under 0.05% of the image power. The origin is shifted to centre the cohort's content box, so voxel coordinates are not MNI coordinates. Intensities are resampled with cubic interpolation and label maps with nearest neighbour. Registered volumes are checked visually. Placement-offset sensitivity is in Appendix C.7.

Evaluation grid and labels. Brain comparisons use the 1.5 mm grid, except the labelled two-grid representation arm. BrLP and the MambaControl backbone use $120\times144\times120$ volumes. SynthSeg (Billot et al., 2023) operates internally at 1 mm, and its labels are returned to the evaluation grid and rounded to integers before mask construction and volume measurement.

J.2 KNEE RADIOGRAPHS

Knee pairs are OAI radiographs (Nevitt et al., 2006; Peterfy et al., 2008) registered within each knee to its earliest visit, split by patient and read in three regions: femur, tibia and the tibiofemoral joint space (JS).

Data. The radiographs are the OAI films of the KXR9 release, $49{,}439$ radiographs of $8{,}295$ knees (Table S37(b)): one 700×700 eight-bit greyscale PNG per knee per visit with a 256×256 integer mask, plus sex, body mass index, age and the Kellgren–Lawrence grade. Elapsed time is read from the nominal OAI schedule, 0, 12, 24, 36, 48, 72 and 96 months for visit codes 0 to 6, rather than from the integer ages. Every radiograph is resampled to the 256×256 mask grid by bilinear interpolation, and all downstream processing uses that grid. No pixel spacing is recorded, so every knee length is in pixels.

Registration. Within each knee the earliest visit is the anchor and every later visit is registered to it in two steps: a centroid-and-scale step on the femur and tibia labels (scale clipped to $[0.92, 1.08]$), then a two-scale Demons deformation in which a band over the tibiofemoral gap, dilated by six pixels, moves rigidly, so the joint space, whose narrowing is the disease signal, cannot be compressed. The Demons step is kept only when the union-label Dice against the anchor falls by no more than 0.01 and the bone centroid moves by no more than half a pixel (87.9% of registered visits); mean union-label Dice against the anchor is 0.951 before registration and 0.961 after it. Images are warped with linear interpolation and masks with nearest neighbour, so a registered visit's stored mask is the release mask carried through the same deformation. Registered radiographs are stored as monotone intensity maps of their originals, anchors as resized, and every method reads the stored radiographs.

The joint space. The release masks label femur, fibula and tibia and leave the gap unlabelled, so the joint space is defined geometrically, as the pixels of each image column strictly between the femur's inferior margin and the tibia's superior margin, and written as a third label beside every stored mask. The fibula is moved to a label the graph ignores; its area change is unrelated to the joint space's (r=0.06 over the strict-QC test pairs). Stray release labels are ignored by the area targets and mapped to background by the evaluation segmenter. Stored-mask and segmenter read-back protocols are in Appendix H.2.

Quality control. Two rules gate the pairs. Registration quality control requires a union-label Dice of 0.90 between the anchor mask and the registered mask (anchors pass); it fails 516 of the $41{,}144$ registered visits (1.3%), and a pair is kept when both visits pass, leaving $25{,}691$ of $26{,}196$ training, $6{,}608$ of $6{,}731$ validation and $8{,}055$ of $8{,}217$ test pairs. The segmentation rule uses the evaluation segmenter, a two-dimensional U-Net of 7.76 M parameters trained on the $31{,}482$ training radiographs against their registered release masks and selected on the $8{,}075$ validation radiographs by mean femur and tibia Dice (0.987 femur, 0.955 fibula, 0.987 tibia at the selected last epoch); a visit is an outlier when the segmenter's fibula on its true radiograph, the label it is least sure of, reads below Dice 0.85 against the stored mask. The rule flags 466 of the $9{,}882$ test visits (4.7%) and 360 of the $8{,}075$ validation visits (4.5%). Of the 700 test pairs the headline table excludes, 92 fail registration quality control alone, 538 the segmentation rule alone and 70 both.

J.3 GRAIN GROWTH

The grain task has exact grain labels: its images are simulated boundary evolution, and its splits are disjoint by microstructure.

Data. Longitudinal images with a fixed field of view and exact grain identities come from the two-dimensional Potts model with Metropolis kinetics (Anderson et al., 1984) in SPPARKS (Plimpton et al., 2009). Each of $1{,}200$ microstructures starts from 150 Voronoi grains on a 256×256 periodic lattice at $kT{=}0.6$ and, after a burn-in of 40 Monte Carlo steps (MCS), is recorded at 0, 100, 200, 300, 450, 600 and 800 steps, over which the grain count falls by nearly half. Samples differ only in the random seed. The image renders the grain map with one fixed grey level per grain, drawn uniformly in $[0.35, 0.9]$, boundaries as a dark line two pixels wide, and Gaussian noise of standard deviation 0.02. Boundary migration changes the intensities of the pixels it crosses; image results are in Table S32 and progression errors in Appendix F. Microstructures are split by sample (Table S37(c)). One hundred MCS stand for a year in the loader's clock, so the previous interval's rate is per hundred steps. There is no other covariate.

Region set. The $K = 32$ nodes are the largest baseline grains that persist to the final snapshot, labelled by size rank. Other grains, including those that vanish, are background. Table S37(c) reports their area-change distribution.

Reading the render. The labeller is a fixed rule, not a trained segmenter. A pixel is a boundary when it is dark (below 0.2), a valley (along some direction both neighbours at the same distance, up to three pixels, are brighter by 0.05 and the pixel is below 0.7) or on a grey step (its two neighbours at distance one differ by more than 0.12); the second and third cues were set on synthetic warps of true images, never on a render. Interior four-connected components are grains, with the periodic wrap honoured, and tracked grain k is the component of at least six pixels that overlaps its baseline footprint, carries its grey level to within 0.03 (read off the baseline image) and is the most contained in that footprint; its area is its cell in the nearest-component partition of the image. The labeller can misidentify a shrinking grain as an expanding neighbour (Appendix F). The rule is applied alike to the true baseline, the render and the true follow-up (read-back ceiling); cells whose reading of a true image is off by more than 25% of the exact area are dropped for every method alike (25 of $24{,}192$ cells on validation and 77 of $49{,}536$ on test). On the true images of either split the rule reads areas to within $\pm 0.5\%$ (interquartile range) and the ceiling is 0.993 (Table S32); on true follow-ups warped by a smooth random field of up to three pixels, as a renderer resamples, the interquartile range widens to -1.0 to $+1.2\%$, so the ceiling read on unwarped truth is a little generous.

K EXTENDED RELATED WORK

K.1 PROGRESSION PREDICTION WITHOUT IMAGE GENERATION

Progression models without image generation forecast diagnoses, biomarkers and clinical outcomes (Young et al., 2024; Marinescu et al., 2019). Event-based and subtype-and-stage models recover biomarker ordering and disease subtypes (Fonteijn et al., 2012; Young et al., 2018). TADPOLE evaluates diagnosis, ADAS-Cog13 and ventricular-volume forecasts (Marinescu et al., 2019). Recurrent models handle irregular visits and missing data (Mehdipour Ghazi et al., 2019; Nguyen et al., 2020), while MC-RVAE models multimodal AD biomarkers with recurrent variational inference (Martí-Juan et al., 2023). DeepAtrophy learns progression-sensitive biomarkers from scan order and interval (Dong et al., 2021). Knee and cardiac models predict structural progression, phenotypes or outcomes (Hu et al., 2023; Lv et al., 2025; Salih et al., 2023; Gao et al., 2023). These methods motivate our tabular reference (Appendix H.2). GraphMorph also predicts spatial anatomy and generates the future image.

K.2 GENERATIVE FORECASTING OF LONGITUDINAL IMAGES

Longitudinal Variational Auto-Encoders couple image generation with temporal mixed-effects modelling (Sauty & Durrleman, 2022). Adversarial models synthesise cardiac remodelling (Campello et al., 2022) and past and future knee radiographs (Prezja et al., 2025). IC-RDN (Wu et al., 2024)

and the diffusion model of Butler et al. (2025) predict 12-month knee radiographs from baseline images. Pix2Pix-style conditional adversarial networks also support paired medical image translation (Isola et al., 2017; Goodfellow et al., 2014). DANI-Net constrains regional intensity progression using regressors trained on age and diagnosis (Ravì et al., 2022).

Rendering a prescribed change. Biophysical simulators deform a scan to match a hand-set atrophy map (Khanal et al., 2017). Xia et al. (2021) synthesise the ageing brain from cross-sectional data conditioned on age and diagnosis. CounterSynth learns a diffeomorphic deformation conditioned on counterfactual age or sex attributes (Pombo et al., 2023). Segmentation-guided diffusion sampling uses anatomical maps (Konz et al., 2024), extending semantic-map conditioning used in natural images (Park et al., 2019). Sun et al. (2025) evaluate counterfactual generators conditioned on regional volumes. GraphMorph's renderer deforms the baseline to a predicted anatomical request.

Counterfactual anatomy and PCGM. PCGM (Li et al., 2025) and GraphMorph both guide image generation with explicit anatomical maps and differ in graph, spatial prediction, supervision and evaluation. *Graph and task.* PCGM's structural causal model relates metadata, the parents, to scalar ROI-volume nodes through conditional normalising flows; an intervention changes metadata while retaining inferred exogenous noise (Li et al., 2025, Section 3.1). In GraphMorph each node is one region's binary segmentation encoded as a spatial tensor; graph-attention layers exchange information between nodes, and the evolver's cross-region attention updates their spatial states over the requested interval, with prior-visit history when available. The forecast is supervised against measured follow-up anatomy, and the attention graph makes no causal-identification claim.

Spatial prediction. PCGM constructs cortical masks by ranking ROI–CSF boundary voxels using segmentation probabilities until the requested volume is reached, preserving the white-matter boundary (their Section 3.2). GraphMorph decodes evolved spatial states into future occupancy maps, learning their geometry under longitudinal supervision. *Rendering and supervision.* PCGM renders by mask-conditioned diffusion with DDIM inversion and regeneration (their Sections 3.3–3.4), reports no explicit loss matching rendered anatomy to its generated masks, and its temporal attention follows diffusion time. GraphMorph trains a control branch against the frozen predicted occupancy maps through overlap and volume losses with the pretrained deformation renderer fixed (Eq. 10), warps the baseline image deterministically at inference, and its evolver advances anatomical states over the inter-visit interval.

Evaluation. PCGM's aging experiment uses CDM without the mask branch; the complete PCGM tests diagnostic counterfactuals through AUD–control group differences after averaging visits within subjects (their Sections 4.1, 4.3–4.4). GraphMorph evaluates subject-specific regional change against observed follow-ups and separately measures request accuracy and rendering adherence.

Concurrent work. Several 2025–2026 preprints address the same problem. AG-LDM supervises a latent diffusion forecaster with a tissue-segmentation loss on ADNI (Wan et al., 2026). TADM-3D extends TADM to 3D with bidirectional temporal regularisation (Litrico et al., 2025), and ProgFormer forecasts in voxel space with conditional flow matching (Ding et al., 2026). TimeFlow extrapolates a temporally conditioned registration (Jian et al., 2025), and MRExtrap extrapolates linearly in an autoencoder latent with subject-specific rates (Kapoor et al., 2025). In GraphMorph, the forecaster outputs replaceable regional occupancy maps that serve as explicit rendering targets (Eq. 10; Section 3.3).

On evaluation, Feng et al. (2026) describe identity collapse, in which optimisation reproduces the current anatomy and misses faint temporal change, and Disch et al. (2025) include the last-context-image predictor, our copy-forward, as a fallback case of their model. Farki et al. (2026) correlate predicted and true change maps at the voxel level. Our change-fidelity protocol is in that tradition: it reads change per named region through one segmenter and frames it with the elapsed-interval, tabular and ceiling references. Re-segmentation measures the resulting regional anatomy (Billot et al., 2023).

Flow matching and neural ODEs. Conditional flow matching learns a velocity field that transports noise to data along a chosen family of conditional probability paths (Lipman et al., 2023). Moschetto et al. (2025) benchmark GANs, diffusion models and flow matching for T1w-to-T2w translation, and ImageFlowNet parameterises multi-scale image trajectories with neural ODEs in-

side a U-Net backbone (Liu et al., 2025; Chen et al., 2018). Three legs of that benchmark, ImageFlowNet and Pix2Pix are in the comparison (Appendix H.1).

Diffusion for longitudinal brain MRI. TADM models the residual change between baseline and follow-up, conditioned on the age gap and baseline features (Litrico et al., 2024). DiffuseMorph instead learns a dense displacement field through a score network and a separate deformation branch, resampling the baseline intensities through that field (Kim et al., 2022); as a registration model, its score network takes the target scan as a conditioning channel, so it can be read as a forecaster only with that channel withheld (Appendix H.5). GraphMorph retains this deformation parameterisation and learns to steer it with predicted anatomy (Section 3.2).

BrLP brings latent diffusion with ControlNet conditioning to 3D brain MRI, optionally with disease-trajectory priors (Puglisi et al., 2025; Rombach et al., 2022; Zhang et al., 2023). Δ-LFM trains a flow-matching model over the change between two visit latents of a 3D autoencoder and scores it with change-aware metrics of its own (Chen et al., 2026). The trajectory it learns is a displacement in that latent, not in any named anatomy. Fu et al. (2025) reach subject-specific trajectories by another route, combining generative modelling with parallel transport from a single baseline MRI.

MambaControl. MambaControl conditions a Mamba latent diffusion backbone on a Fourier-refined graph of image-patch features through a ControlNet (Yang et al., 2025). Its graph is a generator-conditioning feature, whereas GraphMorph decodes its graph into an anatomical request. The comparison uses the released backbone without the graph module (Appendix I.6).

K.3 Deformable registration as a prerequisite

Brain predictions, targets and references share the source pipelines' alignment, with no additional registration (Appendix J). Demons (Thirion, 1998) and its diffeomorphic extensions (Vercauteren et al., 2009) estimate dense displacement fields by intensity matching; longitudinal alignment must correct positional, scale and acquisition differences while preserving pathological change.