LONGITUDINAL MEDICAL WORLD MODELS FOR PATIENT-SPECIFIC DISEASE DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Patients with brain tumours are followed with serial MRI at irregular intervals, and each scan informs whether treatment continues or changes. Anticipating how the tumour will evolve before the next scan could support these decisions. This paper shows that such forecasts are best built from disease dynamics shared across patients, adapted to each patient's treatment and history. We present *Longitudinal Medical World Models* (LMWM). LMWM learns one latent transition from many patients and adapts it with two additive signals: a treatment response field that encodes where and when each recorded course acts, and a history map of the deviations that recur in the patient's own past forecasts. Without these signals, LMWM is exactly the shared model, so the effect of treatment and history can be measured directly. Evaluating such forecasts also needs care: image similarity favours unchanged predictions, and copying the current scan already attains a follow-up lesion Dice of 72.9%. We therefore score the predicted change with a signed change F1 (cF1). On patient timelines built from twelve public brain tumour collections, which align serial MRI with treatment records, LMWM raises cF1 from 31.6% for the shared model to 45.1%, above twelve image-generation baselines, and is the only method whose lesion overlap exceeds that of copying the input. These results point to a simple view of patient-specific forecasting: learn disease dynamics once across patients, adapt them to each individual, and judge the forecast by the change it predicts.

1 INTRODUCTION

Planning cancer treatment and follow-up requires anticipating where and to what extent a patient's tumour will grow or regress over a given interval. The current tumour state provides the starting point for this forecast, while patterns of disease evolution across patients inform its likely course. Treatment can alter that course in ways that vary across space and time, and earlier observations reveal how the tumour has evolved in that patient. Longitudinal tumour forecasting therefore requires models that learn **shared disease dynamics** and adapt them using **treatment information and patient history** to predict individual change.

Patients return for follow-up at irregular intervals, and treatment plans are revised as their condition changes (Figure 1). The same patient may therefore receive different treatments at different times, with each scan reflecting the disease course and preceding interventions. This is the forecasting setting we target: at each visit, predict the next MRI at a specified horizon from the history available so far and a supplied treatment course. Studying this task requires longitudinal data that place scans and treatment records on a common patient timeline, preserving what preceded each forecast and which treatments occurred between visits.

Existing longitudinal image models address this forecasting problem by predicting follow-up scans from the current scan and elapsed time (Liu et al., 2025a; Litrico et al., 2024; Chen et al., 2026). Treatment-conditioned generators additionally incorporate drugs, doses or dose maps (Liu et al., 2025b; Belkacemi et al., 2026; Yang et al., 2025b). These generators feed treatment to a single network as one more input, so how treatment and patient history modify shared disease dynamics, and what each contributes to the forecast, remains unidentified. Evaluation compounds the problem: follow-up lesion overlap rewards not changing the input, and voxelwise change scores reward blurring it. On our test set, copying the current scan already achieves a follow-up lesion Dice of

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

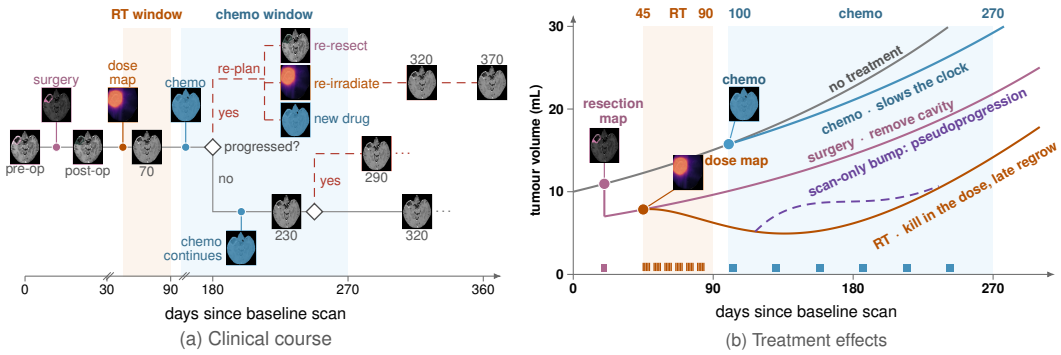

Figure 1: Forecasting the next MRI from patient history and a supplied treatment course. **(a)** Irregular scans and regimen changes define forecast origins. **(b)** Illustrative responses show treatment timing and spatial support; the dashed curve denotes pseudoprogression.

72.9%, and a blurred copy outperforms ten of the twelve baselines on signed change F1 (Section 4). A method can score well while predicting no progression at all.

Structure of longitudinal change in latent space. Figure 2 examines how tumour change relates to disease state, elapsed time and treatment. In the latent space of our change-sensitive autoencoder (Section 2.2), tumour changes agree more across MRI modalities than image differences do and respond less to intensity perturbations (a). Their magnitude is associated with both baseline tumour size and preceding change after accounting for cohort and elapsed time (partial $\rho = 0.69$ and 0.58; b). Changes spanning more shared time also have greater overlap, while their magnitude grows sublinearly with elapsed time (log–log slope 0.10; c). Treatment-aligned trajectories reveal distinct associations: chemotherapy accompanies a lower share of change energy in enhancing tumour, whereas radiotherapy accompanies greater change magnitude (d). These observations motivate a model that predicts change from the current disease state and adapts it using the patient's preceding course and treatment timing. Appendix E gives the analysis protocol and full statistics.

To adapt shared disease dynamics to an individual's treatment and observed course, we propose *Longitudinal Medical World Models* (LMWM; Figure 3). A latent transition, learned across patients, predicts the next disease state from the current one and elapsed time. A *treatment response field* represents where a supplied course acts and how its response evolves over time, and modulates every transformer block. A *consistent map* identifies recurring deviations in the patient's completed forecasts and, together with the record text, enters through a gated attention branch. Both modules adapt the fixed transition additively, preserving shared dynamics while incorporating the individual's care and observed course. With neither treatment nor history inputs, LMWM reduces exactly to the treatment-agnostic transition LMWM°, so the contribution of each module is measured against the same model.

To support forecasting along these evolving patient courses, we harmonize twelve public glioma and brain metastasis collections into one resource of trajectories linking scan intervals, lesion annotations and treatment records (Section 3). For brevity, we refer to the glioma cohort as GBM: glioblastoma accounts for most of its patients, and the remainder have lower-grade glioma. On held-out GBM patients, we compare against twelve image-generation baselines that all receive the same follow-up-time and treatment encodings, and we score the change itself. We introduce signed change precision, recall and F1 (cF1), which score the direction and magnitude of voxelwise change, and add segmentation-derived growth, regression and false growth. Treatment and history conditioning raise cF1 from 31.6% to 45.1% and tumour PSNR from 17.6 to 19.4 dB, above the best baseline on both, and LMWM is the only method whose follow-up lesion Dice exceeds that of copying the input (Section 4).

Our contributions are:

- A longitudinal world model that adapts shared dynamics to treatment and patient history while retaining the treatment-agnostic forecast as an exact special case, so that each adaptation's contribution is measurable.

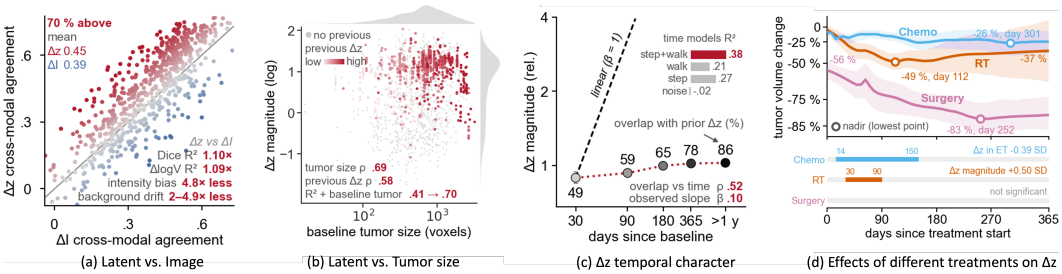

Figure 2: Latent change carries disease-state and temporal structure in GBM. **(a)** Cross-modal agreement; **(b)** associations with tumour size and preceding change; **(c)** temporal scaling and shared-time overlap; **(d)** treatment-aligned volume and latent changes. Volume bands: 68% patient-bootstrap intervals. Definitions and analyses: Appendix E.

- A harmonized resource of twelve brain tumour collections organized into patient timelines, linking serial MRI to treatment records for forecasting across irregular visits and changing treatment courses.
- A change-centred evaluation that exposes what image similarity hides, comparing twelve image-generation baselines on the same test set under identical time and treatment encodings.

Appendix G discusses related work.

2 METHOD

LMWM has three parts (Figure 3): a latent representation of the disease state, to which Section 2.2 adds a change-direction objective; a shared transition F that predicts the next state over the elapsed time (Section 2.3); and two adaptations of the frozen transition, a treatment response field E (Section 2.4) and a consistent map H of the patient's history (Section 2.5). They are fitted in order, each with the earlier ones fixed: representation, transition, treatment field, history branch. Both adaptations are additive and vanish without their inputs: with $E = 0$ and the history branch bypassed, the model is exactly the shared transition. Applied to the current scan alone, with no treatment record and no earlier scan, this transition is the treatment-agnostic forecast LMWM° (Section 4). Appendix C gives the architecture and training details.

2.1 PREDICTION SETTING

For patient p, let $x_{p,i}$, $s_{p,i}$ and $t_{p,i}$ denote the registered image, segmentation and acquisition time, and let c_p denote patient covariates (age and sex, with presence indicators). The observed prefix is

$$\mathcal{P}_{p,i} = \{(x_{p,j}, s_{p,j}, t_{p,j}) : j \le i\} \cup \{c_p, \text{records available by } t_i\}. \tag{1}$$

Given $\mathcal{P}_{p,i}$, horizon $\Delta_i > 0$ and a supplied treatment course a_i for the interval, the model predicts the next image; a predicted lesion mask gates its readout. In evaluation, a_i is the recorded course for the interval (Appendix C.5). The course specifies treatment class, timing and available spatial support (Figure 1); missing fields retain availability indicators. We omit p below.

Target images, target segmentations and endpoint-derived treatment supports provide supervision only. We train and evaluate the forecasting model on the GBM cohort. All observations from one person share a split. Forecasting pairs require a verified positive interval between scans, whereas representation learning needs only a reliable visit order. Appendix C.5 specifies the information regimes.

2.2 LEARNING A CHANGE-SENSITIVE REPRESENTATION

Let $z_i = \mathrm{Enc}_\psi(x_i)$ be the encoder's posterior mean and D_ω its decoder. Reconstruction objectives leave latent differences unconstrained, so we add an objective that makes the change predictable. For three consecutive visits, an auxiliary affine predictor estimates $d_i = z_{i+1} - z_i$ from $[z_i,\ z_i - z_{i-1}]$. The loss is one minus its channel-averaged cosine agreement with d_i, plus the agreement between

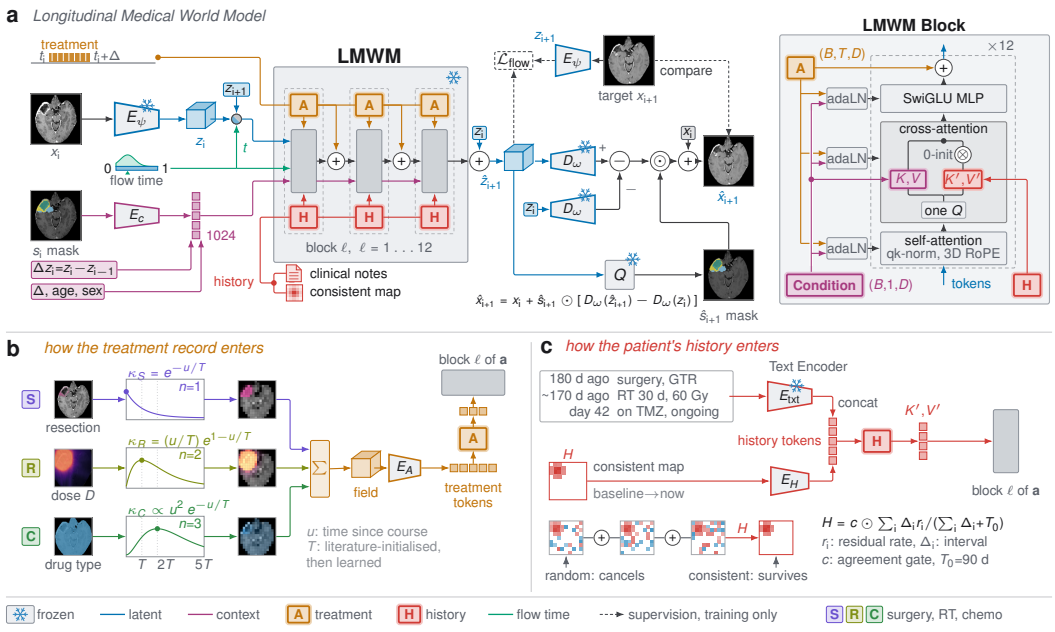

Figure 3: **LMWM adapts shared dynamics to treatment and patient history. (a)** A latent transition predicts the next state; decoded change is added within the predicted lesion. Inset: treatment modulates normalization and adds a block-output residual; history adds keys and values through gated attention. **(b)** Spatial treatment codes and course calendars form response fields through learned transit kernels. **(c)** Record text and a consistent map of historical prediction errors supply history tokens.

a detached running mean change and d_i, so the objective rewards patient-specific change rather than the average drift of the cohort. Whole-grid and tumour-region terms are added with a ramped weight to the reconstruction, perceptual, adversarial and change-reconstruction losses. Appendix F explains why the objective supervises change direction rather than the next state. Appendix C.3 gives the loss, masks and normalization. The objective makes the autoencoder preserve more of the observed change: the cF1 between the residual of two acquired scans and the residual of their reconstructions, computed over the whole brain and averaged over T1 and T1ce, is 82% for the public MAISI autoencoder, 87% after reconstruction fine-tuning and 89% with the change objective. After training, the predictor is discarded and the encoder and decoder are fixed.

2.3 Shared Latent Transition

The shared transition carries the current state to the endpoint of the interval. It is trained on straight latent paths by flow matching (Lipman et al., 2023; Liu et al., 2023) (Figure 3a):

$$\hat{z}_{i+1} = F_\theta(z_i; \phi_i) = z_i + \int_0^1 v_\theta\big(z^t, t; \phi_i\big)\, dt, \qquad \phi_i = (\Delta_i,\, c_p,\, s_i,\, z_i - z_{i-1}), \tag{2}$$

integrated from $z^0 = z_i$ with five Euler steps. The velocity field v_θ is a twelve-block transformer over the 4,096 tokens formed from 2^3-voxel latent patches (Peebles & Xie, 2023). The flow time t, horizon Δ_i and covariates modulate every block and the output normalization through a broadcast condition. The current segmentation and the preceding latent change supply the cross-attention memory, with a learned missing-history token for LMWM$^\circ$, LMWM† and queries without an earlier scan. The flow loss $\mathcal{L}_{\text{flow}}$ is an L1 penalty between the predicted and observed residual $z_{i+1} - z_i$, together with a latent adversarial term of weight 0.1 and the Dice loss of the lesion readout (Section 2.6, Appendix C.10). The transition and its condition encoders are fixed before adaptation.

2.4 Treatment-Conditioned Adaptation

Treatment response depends on both where a course acts and when it is delivered (Figure 3b). For treatment class k, an encoder N_k maps current anatomy, the source plan and plan attributes p_k

to a spatial code $\vartheta_k(x)$. For systemic therapy, p_C includes a one-hot drug/regimen category, so the spatial response depends on both treatment identity and patient anatomy. Following transit-compartment models of delayed anticancer drug effects (Sun & Jusko, 1998; Simeoni et al., 2004), each response function h_k combines a class-specific kernel, with an immediate or delayed peak, and a learned correction; its time constants are initialized by treatment class and learned from $\vartheta_k(x)$ and p_k. Calendar weights w_{ikg} summarize the course's timing over the forecast interval:

$$E_i(x) = \sum_{k \in \mathcal{K}_i} \sum_g w_{ikg}\, h_k(u_g; \vartheta_k(x), p_k)\,. \tag{3}$$

Here u_g indexes time since treatment and $\mathcal{K}_i$ contains the recorded classes. Spatial features determine the response profile, while the calendar determines its contribution to this forecast. Repeated and concurrent courses add before the transformer processes their combined field.

A projector E_A, pooling followed by an MLP, maps E_i to the token space, and shared low-rank dictionaries project the result into two blockwise adaptations: scale and shift corrections to each sub-block's adaptive normalization, and a residual of the block's width added after the final feed-forward sub-block. The first changes how attention and the feed-forward layer read the state; the second updates the block output directly. Both projections are bias-free, so without an eligible dated record $E_i = 0$ and every block reproduces the shared transition exactly. Kernel, calendar and block equations appear in Appendix C.7.

2.5 Patient history through consistent residuals

A patient's past is informative when it reveals deviations that recur beyond the treatment-conditioned prediction. Let F_E denote the treatment-conditioned transition. For each completed interval $j < i$, define the normalized residual rate

$$r_j = \frac{z_{j+1} - F_E(z_j; \phi_j, E_j)}{\Delta_j\, \sigma_r}, \tag{4}$$

the observed endpoint minus the treatment-conditioned forecast, per day and divided by $\sigma_r(x)$, the voxelwise standard deviation of that per-day error over cross-fitted (fold-held-out) training intervals, pooled over channels, floored at 10^{-6} and estimated once; r_j is dimensionless. The *consistent map* combines duration weighting, shrinkage for short histories and directional coherence:

$$H_i = c_i \odot \frac{\sum_{j<i} \Delta_j r_j}{\sum_{j<i} \Delta_j + T_0}, \qquad c_i(x) = \frac{\| \sum_{j<i} \Delta_j r_j(x) \|_2}{\sum_{j<i} \Delta_j \| r_j(x) \|_2 + \epsilon}. \tag{5}$$

Here $T_0 = 90$ days, $\epsilon > 0$, and norms are over channels. Aligned residuals accumulate; conflicting directions reduce the coherence weight. Empty imaging history gives $H_i = 0$. Appendix C.9 derives the weighted estimator and explains the coherence factor.

A projector E_H, pooling followed by an MLP, maps H_i to spatial tokens, which are concatenated with text tokens of the patient's record (dates, extents, regimens) from a fixed clinical text encoder. The text tokens carry record content that the field does not encode, such as resection extent or regimen names. Each block projects these tokens into history keys and values, K'_ℓ, V'_ℓ, and reuses its cross-attention query $\mathbf{q}$ (per head, width d) in a separately normalized attention branch:

$$\mathrm{out}_\ell = \mathrm{softmax}\big(\mathbf{q}K_\ell^\top / \sqrt{d}\big)V_\ell + \tanh(\lambda_\ell)\, \mathrm{softmax}\big(\mathbf{q}K_\ell'^\top / \sqrt{d}\big)V'_\ell. \tag{6}$$

This branch has its own softmax, so history tokens do not compete with the context tokens for the same query (Alayrac et al., 2022). Each layer's scalar gate λ_ℓ is initialized to zero, so training starts exactly at the treated transition; the gate still receives gradient because $\tanh$ has unit slope at zero. History projectors, gates and the map encoder are learned while the treated transition remains fixed. The branch shares the block's output projection and residual gate. With no map and no text, it is bypassed and the treated transition is recovered exactly; $H_i = 0$ alone leaves any available text active.

2.6 Training and image prediction

Treatment adaptation learns the spatial encoders, response functions, projector E_A and dictionaries on the fixed transition, giving F_E. History adaptation fixes F_E and learns the projector E_H, the token

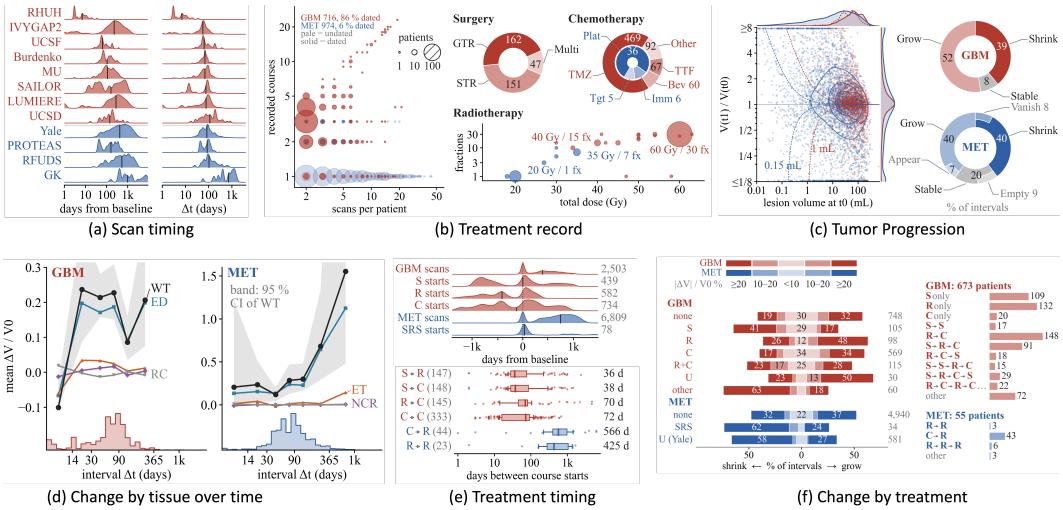

Figure 4: Longitudinal sampling and recorded treatment vary across cohorts, with both tumour growth and regression. **(a)** Scan timing; **(b)** treatment records and imaging depth; **(c)** baseline size and volume change; **(d)** tissue changes by scan interval; **(e)** treatment timing and gaps between course starts; **(f)** change by recorded treatment and treatment sequences. Red: glioma; blue: metastasis. Whole-tumour bands: 95% patient-bootstrap intervals. Populations and definitions: Appendix D.

projection to the context width, the history key–value projections and the gates, using only intervals completed before each query. Splits and objectives are specified in Appendices C.10 and C.11.

A readout head Q predicts the follow-up lesion mask $\hat{s}_{i+1} = Q(\hat{z}_{i+1})$ from the forecast latent; it is trained with the transition by a Dice loss against the follow-up segmentation. The image prediction adds decoded change to the current scan:

$$\hat{x}_{i+1} = x_i + \hat{s}_{i+1} \odot [D_\omega(\hat{z}_{i+1}) - D_\omega(z_i)] \,. \tag{7}$$

This keeps the acquired image outside the predicted lesion and confines the forecast to the tumour. All LMWM variants, including LMWM°, use this readout. Grid details appear in Appendix C.12.

3 Longitudinal Disease Progression Dataset

We harmonize twelve public brain tumour collections (eight glioma and four metastasis sources) into patient timelines that place each scan within the preceding course and the treatments delivered between visits, linking MRI, lesion annotations, scan times and treatment records. Table S4 lists source references, links and licenses; Appendix D.2 defines eligibility. The split inventories contain 816 glioma and 1,029 metastasis records. After quality control, 800 glioma records with 2,613 visits and 1,000 metastasis records with 6,840 visits remain. Of these, 273 glioma and 775 metastasis records have at least three visits, allowing a completed interval to precede the target visit. Splits group all records of the same person.

Trajectory structure. Scan intervals and treatment timing vary across collections (Figure 4a, e). Median intervals are 68 days for glioma and 85 days for metastases; 59% and 83% of included pairs have earlier scans available (Table S5). Both growth and regression occur within treatment groups. Changes occupy a median 2.3% and 0.5% of brain voxels, respectively (Table S6), motivating forecasts that resolve local change within a patient's ongoing course.

Processing and task eligibility. Scans from all visits of each patient are registered to one $192 \times 192 \times 160$ grid at 1 mm resolution. Quality control is performed within and across visits (Appendix D). We retain expert segmentations supplied with the original datasets and use automated segmentation when annotations are unavailable. Missing MRI sequences are generated using a modality-synthesis model (Appendix D). Provenance records distinguish acquired from synthesized sequences and expert from generated masks, with treatment availability recorded separately.

Table 1: Treatment ($+A$) and history ($+H$) improve tumour-region cF1 and PSNR; perceptual distance favours near-copy outputs.

		Appearance		Change agreement (%)			
Method	Venue	PSNR (dB) ↑	LPIPS $\times 10^3$ ↓	cF1 ↑	cPrec ↑	cRec ↑	ρ ↑
Copy input	–	16.97±4.83	2.75±3.97	0.0	0.0	0.0	0.0
Blur input	–	–	–	32.7±14.6	–	–	40.6±16.5
Δ-LFM	ICLR'26	17.43±3.54	8.00±8.15	40.6±21.4	60.8±30.5	36.0±22.5	43.4±21.3
MambaControl	arXiv'25	17.31±3.17	16.73±13.47	41.4±20.5	61.8±33.3	36.8±19.3	45.0±22.5
ImageFlowNet	ICASSP'25	17.71±4.21	4.89±5.21	28.1±16.2	**73.1**±34.2	19.3±12.4	42.0±22.2
GlioPREDICT	arXiv'26	16.27±3.33	3.60±4.25	27.0±19.0	47.2±33.2	23.8±18.6	28.9±18.2
CardiacAging	Front.CVM'22	17.02±3.95	3.00±3.99	21.3±16.3	68.2±36.1	14.5±12.3	28.1±15.9
CounterSynth	MedIA'23	16.82±4.14	3.00±4.14	19.0±15.8	53.9±34.9	13.4±12.0	29.6±16.1
DiffuseMorph	ECCV'22	16.96±4.31	**2.86**±3.99	6.9±6.8	48.7±35.0	7.4±6.5	15.6±10.6
MeWM	ICCV'25	16.67±3.97	2.99±3.93	16.2±11.5	49.3±30.5	11.3±9.1	22.7±13.0
BrLP	MICCAI'24	15.95±3.61	3.29±4.09	15.8±11.9	34.2±26.8	13.4±12.8	18.6±14.5
TADM	MICCAI'24	16.48±4.07	2.89±3.80	13.4±11.4	43.3±30.4	9.8±10.2	15.0±12.0
SADM	IPMI'23	14.97±2.97	3.70±3.99	17.5±9.3	30.4±20.8	15.7±8.9	18.8±10.6
TaDiff	TMI'25	15.81±3.55	2.93±3.40	19.8±10.4	38.8±23.7	15.4±9.2	22.3±9.2
LMWM°	Ours (Base)	17.63±3.81	5.41±7.26	31.6±19.7	67.4±32.1	22.9±16.4	39.4±21.6
LMWM†	Ours (+A)	18.29±3.53	4.94±6.58	37.2±19.9	63.3±29.2	35.4±20.6	43.7±20.9
LMWM	Ours (+A+H)	**19.38**±2.67	3.66±4.63	**45.1**±18.6	63.8±30.3	**39.3**±21.7	**49.7**±21.6

Mean ± case SD in the tumour region; bold: best among learned methods. –: the blur reference serves only as a change-score floor. Base: current scan only; A: treatment field E; H: preceding latent change, consistent map and record text. Definitions: Appendix B.

Ordered visits support representation learning, and verified positive intervals support forecasting. Patient adaptation also requires completed intervals before the target visit. Endpoint-derived cavity maps are outcomes, whereas eligible treatment inputs use independently sourced records and plans. Appendices D and C.5 detail the pipeline and timing rules.

4 Results

Evaluation setting. We evaluate forecasts on the GBM cohort with an 80/10/10 patient split. Each adjacent scan pair and sequence (T1 or contrast-enhanced T1) is a case; all methods share the test cases. Within the union of current and follow-up tumour masks, we introduce signed change precision (cPrec), recall (cRec) and their harmonic mean, cF1. With observed change d and predicted change $\hat{d}$, $\text{cRec} = \text{clip}_{[0,1]}(\langle \hat{d}, d\rangle / \|d\|^2)$ and $\text{cPrec} = \text{clip}_{[0,1]}(\langle \hat{d}, d\rangle / \|\hat{d}\|^2)$, each over voxels with substantial change. Copying or reversing change scores zero. Table 1 also reports tumour PSNR, 3D perceptual distance and voxelwise Spearman correlation. Table S1 uses one fixed segmenter to assess overlap, growth ranking, volume concordance, direction and false growth. Both tables include two references: copying the current scan and a 5^3 box-filtered copy of it. Appendix B defines all scores and aggregation rules.

Matched conditioning. All twelve baselines receive the same horizon and treatment-plan information: time is encoded as in Δ-LFM (Chen et al., 2026), drugs as in MeWM (Yang et al., 2025b) and spatial plans as in GlioPREDICT (Belkacemi et al., 2026). LMWM° receives the current scan, segmentation, horizon and covariates, with no treatment record or earlier scan. LMWM† adds the treatment field, with the preceding-change input held at its missing-history token; LMWM adds the preceding latent change, consistent map and record text (Figure 3). These variants share the representation and readout (Eq. 7). Treatment-conditioned forecasts use the recorded interval course. Δ-LFM is retrained with the same autoencoder as LMWM; other baselines retain their original representations.

Treatment and history recover more of the observed change. LMWM has the highest tumour PSNR (19.38 dB), cF1 (45.1%), change recall (39.3%) and rank correlation (49.7%) in Table 1.

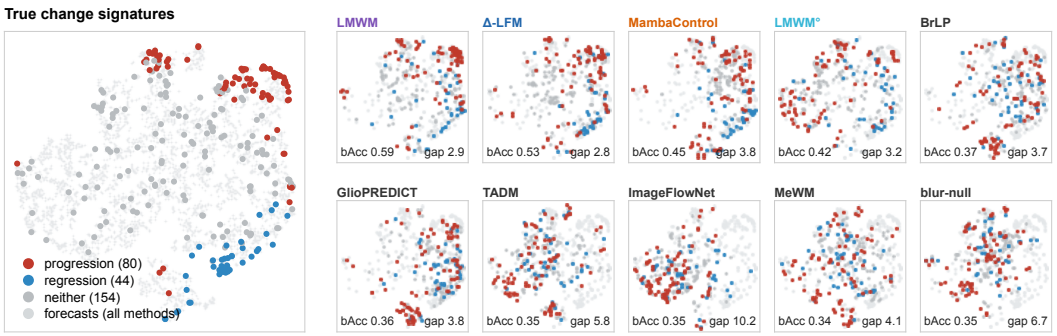

Figure 5: Outcome alignment and coverage assess different aspects of spatial change. Left: observed changes; right: forecasts in the same t-SNE embedding. Colours denote observed progression, regression or neither. Higher neighbourhood bAcc indicates better outcome alignment in nine-dimensional signature space; lower gap indicates closer coverage of observed changes in the embedding.

Both adaptations contribute. The treatment response field raises cF1 from 31.6% to 37.2% and PSNR from 17.63 to 18.29 dB, already above every baseline in PSNR; the history inputs raise cF1 further to 45.1%, PSNR to 19.38 dB and ρ from 43.7% to 49.7%. Most of the cF1 gain comes from recall: cRec rises from 22.9% to 35.4% and 39.3%, while cPrec remains between 63.3% and 67.4%. The strongest baselines reach cF1 41.4% (MambaControl) and 40.6% (Δ-LFM). ImageFlowNet has the highest cPrec (73.1%) but low cRec (19.3%): its forecasts capture only a small share of the observed change.

Lesion overlap conceals change errors. Copying the input attains a follow-up lesion Dice of 72.9% in Table S1, above every baseline, including DiffuseMorph (72.6%), CounterSynth (72.3%) and ImageFlowNet (71.1%), whose volume-change concordances are -0.4%, -0.6% and 4.2%. LMWM is the only method that exceeds it (75.7%) while also predicting change. A blurred input reaches cF1 32.7% and ρ 40.6%, above ten of the twelve baselines on cF1 and nine on ρ. The copy and blur references score high on lesion overlap and voxelwise agreement without any learned forecast, so they set the reference level for every change score.

History sharpens growth prediction. In Table S1, LMWM ranks growth best (AUC 76.2%) and leads in volume-change concordance (47.6%), core growth Dice (18.3%) and direction accuracy (55.0%). History makes the model commit to growth: from LMWM† to LMWM, growth AUC rises from 68.8% to 76.2% and concordance from 40.7% to 47.6%, at a false-growth rate of 67%, below that of the two strongest growth-ranking baselines, Δ-LFM (76%, AUC 71.5%) and MambaControl (97%, AUC 64.7%). LMWM° matches MambaControl in growth AUC (64.9% versus 64.7%) at 38% versus 97% false growth. Copying the input has zero false growth because it predicts no change.

Recovering the distribution of spatial change. Figure 5 summarizes tissue loss and gain by distance from the baseline tumour boundary. LMWM has the highest neighbourhood balanced accuracy (0.59) and a coverage gap of 2.9, level with the best baseline (Δ-LFM, 2.8). ImageFlowNet concentrates its forecasts in a narrow region, with near-chance balanced accuracy (0.35) and a larger gap (10.2). Signature construction and scoring are defined in Appendix A.2.

Where the forecasts differ. Figure 6 distinguishes three behaviours. CounterSynth, ImageFlowNet, MeWM, BrLP and CardiacAging largely preserve the input, including the enhancing focus that disappears at follow-up in the upper case. Δ-LFM, MambaControl and LMWM° produce more diffuse, blurred changes. LMWM changes the lesion while preserving anatomical detail; committing to change also carries risk, and in the upper case it retains part of that focus. This commitment is what gives LMWM the highest change recall in Table 1. Figure S2 provides additional cases.

Appendix A analyses change detection, volume agreement and spatial coverage.

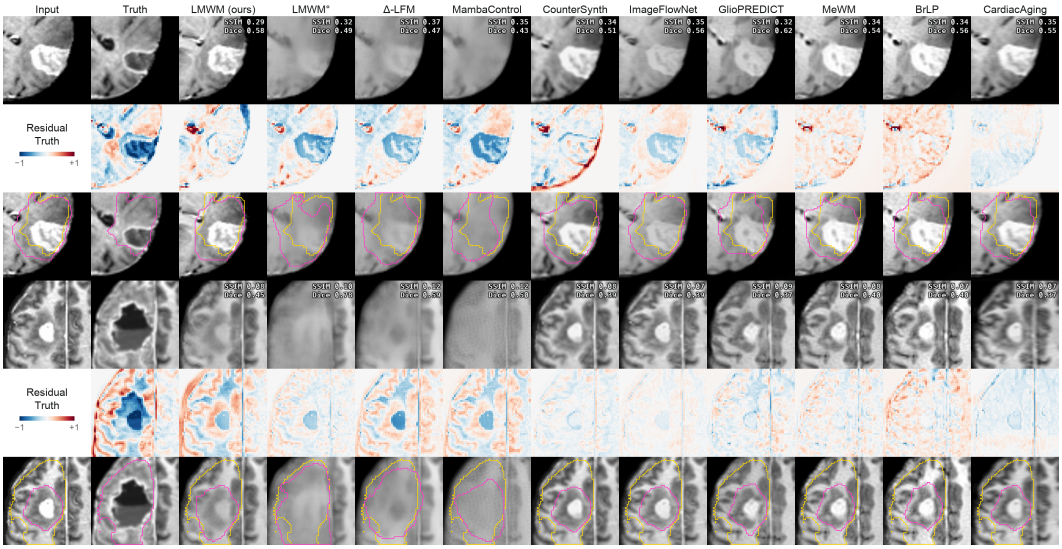

Figure 6: **Baselines either preserve the input or blur the change; LMWM changes the lesion while preserving anatomy.** Rows: images, baseline-relative residuals (red: brighter; blue: darker) and tumour contours. Truth shows observed change. One fixed segmenter supplies both contours: yellow for the acquired follow-up and magenta for the displayed image (baseline in Input). Labels: tumour-region 2D SSIM and 3D Dice.

5 Conclusion

We introduced *Longitudinal Medical World Models* (LMWM), which forecast a patient's disease by learning dynamics shared across patients and adapting them to the individual's treatment and history. Because both adaptations vanish without their inputs, the shared forecast remains an exact reference, and the contribution of each adaptation can be read directly: on GBM, treatment and history together raise signed change F1 from 31.6% to 45.1%, beyond every baseline. Our evaluation also shows that image similarity rewards forecasts that leave the tumour unchanged; scoring the predicted change separates models that forecast from models that copy.

Limitations

Forecasting is evaluated on the GBM cohort, where public treatment records are often incomplete. Few cases include radiotherapy dose maps, resection maps, complete systemic regimens, and treatment dates. Because treatment inputs describe observed care, treatment-conditioned improvements reflect predictive associations, not causal effects. LMWM° is trained on treated patients without treatment records and therefore represents expected disease courses under care received in the training cohorts, not untreated tumour growth.

Where expert annotations or sequences are missing, the resource uses automated segmentation and modality synthesis, and provenance flags mark every generated mask and synthesized sequence.

AI USE STATEMENT

We use Claude and Codex to assist manuscript drafting and editing, and code writing. The scientific ideas, research hypotheses, method and experimental design were developed by the authors. The authors take responsibility for the final manuscript, code and analyses.

ETHICS STATEMENT

This study is a retrospective analysis of twelve existing de-identified brain-tumour imaging collections, cited in Appendix D.2. It involves no new patient recruitment, data collection or clinical intervention. Data use and sharing follow the licenses and access conditions of the source collections. Public releases of processed images and derived annotations will preserve de-identification and respect each source's redistribution terms; data subject to access restrictions will remain available through the original providers.

REPRODUCIBILITY STATEMENT

We will publicly release the code for preprocessing, quality control, model implementation and evaluation, together with configurations, split manifests and instructions for reproducing the reported analyses. Processed data and derived annotations will be shared under the source collections' terms, with processing scripts and source mappings provided for data that require access through the original repositories. Appendices C, B, E and D describe implementation, metrics, representation analyses and data curation, respectively.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop (ClinicalNLP@NAACL)*, pp. 72–78, 2019.

Brian B. Avants, Nicholas J. Tustison, Gang Song, Philip A. Cook, Arno Klein, and James C. Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3):2033–2044, Feb 2011. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2010.09.025. URL http://dx.doi.org/10.1016/j.neuroimage.2010.09.025.

Nordin Belkacemi, Selena Huisman, Vera C. Keil, Joost J. C. Verhoeff, and Szabolcs David. 3D-GlioPREDICT: 3D Latent Diffusion for Post-Radiotherapy Brain MRI Prediction in Patients with Glioma, 2026. URL https://arxiv.org/abs/2606.05113.

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, Sept 2022. ISSN 2297-055X. doi: 10.3389/fcvm.2022.983091. URL http://dx.doi.org/10.3389/fcvm.2022.983091.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. MONAI: An open-source framework for deep learning in healthcare, 2022. URL https://arxiv.org/abs/2211.02701.

Santiago Cepeda, Sergio García-García, Ignacio Arrese, Francisco Herrero, Trinidad Escudero, Tomás Zamora, and Rosario Sarabia. The Río Hortega University Hospital Glioblastoma dataset: A comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM). *Data in Brief*, 50:109617, Oct 2023a. ISSN 2352-3409. doi: 10.1016/j.dib.2023.109617. URL http://dx.doi.org/10.1016/j.dib.2023.109617.

Santiago Cepeda, Sergio García García, Ignacio Arrese, Francisco Herrero, Trinidad Escudero Caro, Tomás Zamora Martínez, and Rosario Sarabia. The Río Hortega University Hospital Glioblastoma dataset: a comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM), 2023b. URL https://www.cancerimagingarchive.net/collection/rhuh-gbm/. Dataset. https://doi.org/10.7937/4545-C905.

Saahil Chadha, David Weiss, Anastasia Janas, Divya Ramakrishnan, Thomas Hager, Klara Osenberg, Klara Willms, Joshua Zhu, veronica chiang, Spyridon Bakas, Nazanin Maleki, Durga V. Sritharan, Sven Schoenherr, Malte Westerhoff, Matthew Zawalich, Melissa Davis, Ajay Malhotra, Khaled Bousabarah, Cornelius Deuschl, MingDe Lin, Sanjay Aneja, and Mariam Aboian. Yale longitudinal dataset of brain metastases on MRI with associated clinical data (Yale-Brain-Mets-Longitudinal), 2025a. URL https://www.cancerimagingarchive.net/collection/Yale-Brain-Mets-Longitudinal. Dataset. https://doi.org/10.7937/3YAT-E768.

Saahil Chadha, David Weiss, Anastasia Janas, Divya Ramakrishnan, Thomas Hager, Klara Osenberg, Klara Willms, Joshua Zhu, Veronica Chiang, Spyridon Bakas, Nazanin Maleki, Durga V. Sritharan, Sven Schoenherr, Malte Westerhoff, Matthew Zawalich, Melissa Davis, Ajay Malhotra, Khaled Bousabarah, Cornelius Deuschl, MingDe Lin, Sanjay Aneja, and Mariam S. Aboian. An 11,000-Study Open-Access Dataset of Longitudinal Magnetic Resonance Images of Brain Metastases, 2025b. URL https://arxiv.org/abs/2506.14021.

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 C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 82707–82723, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/859c3ab5fe31e8d1c1a70a34fe03dde9-Paper-Conference.pdf.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer Learning for 3D Medical Image Analysis, 2019. URL https://arxiv.org/abs/1904.00625.

Yaseen Dhemesh, Filip Garrett, Jaime Gass, Josh Greaser, Edvin Isufi, Lester J. Layfield, Ayman Nada, Karolina Pogorzelski, Jason Sinclair, Nourel Hoda M. Tahon, and Jonathan Thacker. University of Missouri Post-operative Glioma Dataset, 2025. URL https://www.cancerimagingarchive.net/collection/mu-glioma-post/. Dataset. https://doi.org/10.7937/7K9K-3C83.

Margarida Fernandes, José Soares, Matheus Silva, Crystian Saraiva, and Victor Alves. RFUDS – A Brain Metastases Imaging Dataset of Radiotherapy Follow-Up, 2024. URL https://arxiv.org/abs/2412.16568.

Brandon K. K. Fields, Evan Calabrese, John Mongan, Soonmee Cha, Christopher P. Hess, Leo P. Sugrue, Susan M. Chang, Tracy L. Luks, Javier E. Villanueva-Meyer, Andreas M. Rauschecker, and Jeffrey D. Rudie. The University of California San Francisco Adult Longitudinal Post-Treatment Diffuse Glioma MRI Dataset. *Radiology: Artificial Intelligence*, 6(4):e230182, 2024. doi: 10.1148/ryai.230182.

Dimitra Flouri, Christos-Panagiotis Papanikas, Georgios C. Manikis, Eleftherios Ioannou, Georgia Karaoli, Elisavet Papageorgiou, Anastasia Constantinidou, Loizos Siakallis, Marilena Theodorou, and Vasileios Vavourakis. A longitudinal MRI dataset of brain metastases with tumor segmentations, clinical & radiomic data. *Scientific Data*, 12(1):1828, Nov 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-06131-0. URL https://doi.org/10.1038/s41597-025-06131-0.

Louis Gagnon, Diviya Gupta, George Mastorakos, Nathan White, Vanessa Goodwill, Carrie McDonald, Thomas Beaumont, Christopher Conlin, Tyler Seibert, Uyen Nguyen, Jona Hattangadi-Gluth, Santosh Kesari, Jessica Schulte, David Piccioni, Kathleen Schmainda, Nikdokht Farid, Anders Dale, and Jeffrey Rudie. The University of California San Diego annotated post-treatment high-grade glioma multimodal MRI dataset (UCSD-PTGBM), 2025. URL https://www.cancerimagingarchive.net/collection/ucsd-ptgbm/. Dataset. https://doi.org/10.7937/FWV2-DT74.

Louis Gagnon, Diviya Gupta, Uyen Nguyen, Maria Correia de Verdier, Rachit Saluja, George Mastorakos, Nathan White, Vanessa Goodwill, Carrie R. McDonald, Thomas Beaumont, Christopher Conlin, Tyler M. Seibert, Jona Hattangadi-Gluth, Santosh Kesari, Jessica D. Schulte, David Piccioni, Kathleen M. Schmainda, Nikdokht Farid, Anders M. Dale, and Jeffrey D. Rudie. The University of California San Diego Post-Treatment Glioblastoma (UCSD-PTGBM) annotated multimodal MRI Dataset. *Scientific Data*, 13(1):183, Jan 2026. ISSN 2052-4463. doi: 10.1038/s41597-025-06499-z. URL https://doi.org/10.1038/s41597-025-06499-z.

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 *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 4430–4441, February 2025.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/2de5d16682c3c35007e4e92982f1a2ba-Paper.pdf.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, Apr 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-08744-2. URL http://dx.doi.org/10.1038/s41586-025-08744-2.

James Harrison, Apoorva Sharma, and Marco Pavone. Meta-Learning Priors for Efficient Online Bayesian Regression. In *Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2018. URL https://arxiv.org/abs/1807.08912.

Andrew Hoopes, Jocelyn S. Mora, Adrian V. Dalca, Bruce Fischl, and Malte Hoffmann. SynthStrip: skull-stripping for any brain image. *NeuroImage*, 260:119474, Oct 2022. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2022.119474. URL http://dx.doi.org/10.1016/j.neuroimage.2022.119474.

Ivar T. Hovden, Elies Fuster-Garcia, Christopher Larsson, Jonas Vardal, Magne M. Kleppestø, Torstein R. Meling, Petter Brandal, Inge R. Groote, and Atle Bjørnerud. Brain tumor MRI: Serial assessments in longitudinal oncological research (SAILOR) (v1), 2023. URL https://search.kg.ebrains.eu/instances/cae85bcb-8526-442d-b0d8-a866425efff8. Dataset. https://doi.org/10.25493/D948-J3P.

Boah Kim, Inhwa Han, and Jong Chul Ye. DiffuseMorph: Unsupervised Deformable Image Registration Using Diffusion Model. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 347–364, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19821-2.

Wojciech Krzyzanski. Interpretation of transit compartments pharmacodynamic models as lifespan based indirect response models. *Journal of Pharmacokinetics and Pharmacodynamics*, 38(2):179–204, Apr 2011. ISSN 1573-8744. doi: 10.1007/s10928-010-9183-z. URL https://doi.org/10.1007/s10928-010-9183-z.

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.

Mattia Litrico, Francesco Guarnera, Mario Valerio Giuffrida, Daniele Ravì, and Sebastiano Battiato. TADM: Temporally-Aware Diffusion Model for Neurodegenerative Progression on Brain MRI. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 444–453, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72069-7.

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 *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, Apr 2025a. doi: 10.1109/icassp49660.2025.10890535. URL http://dx.doi.org/10.1109/icassp49660.2025.10890535.

Qinghui Liu, Elies Fuster-Garcia, Ivar Thokle Hovden, Bradley J. MacIntosh, Edvard O. S. Grødem, Petter Brandal, Carles Lopez-Mateu, Donatas Sederevičius, Karoline Skogen, Till Schellhorn,

Atle Bjørnerud, and Kyrre Eeg Emblem. Treatment-Aware Diffusion Probabilistic Model for Longitudinal MRI Generation and Diffuse Glioma Growth Prediction. *IEEE Transactions on Medical Imaging*, 44(6):2449–2462, June 2025b. ISSN 1558-254X. doi: 10.1109/tmi.2025.3533038. URL http://dx.doi.org/10.1109/tmi.2025.3533038.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Esmat Mahmoud, Jaime Gass, Yaseen Dhemesh, Josh Greaser, Karolina Pogorzelski, Edvin Isufi, Filip Garrett, Jonathan Thacker, Nourel Hoda Tahon, Jason Sinclair, Lester Layfield, Talissa Altes, and Ayman Nada. MU-Glioma Post: A comprehensive dataset of automated MR multi-sequence segmentation and clinical features. *Scientific Data*, 12(1):1847, Nov 2025. ISSN 2052-4463. doi: 10.1038/s41597-025-06011-7. URL https://doi.org/10.1038/s41597-025-06011-7.

Xudong Mao, Qing Li, Haoran Xie, Raymond Y.K. Lau, Zhen Wang, and Stephen Paul Smolley. Least Squares Generative Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

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, Feb 2023. ISSN 1361-8415. doi: 10.1016/j.media.2022.102723. URL http://dx.doi.org/10.1016/j.media.2022.102723.

Lemuel Puglisi, Daniel C Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 173–183. Springer, 2024.

Jugal K Shah, Matthew B Potts, Penny K Sneed, Manish K Aghi, and Michael W. McDermott. Surgical Cavity Constriction and Local Progression Between Resection and Adjuvant Radiosurgery for Brain Metastases. *Cureus*, Apr 2016a. ISSN 2168-8184. doi: 10.7759/cureus.575. URL http://dx.doi.org/10.7759/cureus.575.

Nameeta Shah, Xu Feng, Michael Lankerovich, Ralph B. Puchalski, and Bart Keogh. Data from Ivy GAP, 2016b. URL https://www.cancerimagingarchive.net/collection/ivygap/. Dataset. https://doi.org/10.7937/K9/TCIA.2016.XLWAN6NL.

Monica Simeoni, Paolo Magni, Cristiano Cammia, Giuseppe De Nicolao, Valter Croci, Enrico Pesenti, Massimiliano Germani, Italo Poggesi, and Maurizio Rocchetti. Predictive Pharmacokinetic-Pharmacodynamic Modeling of Tumor Growth Kinetics in Xenograft Models after Administration of Anticancer Agents. *Cancer Research*, 64(3):1094–1101, Feb 2004. ISSN 1538-7445. doi: 10.1158/0008-5472.can-03-2524. URL http://dx.doi.org/10.1158/0008-5472.CAN-03-2524.

Yu-Nien Sun and William J. Jusko. Transit compartments versus gamma distribution function to model signal transduction processes in pharmacodynamics. *Journal of Pharmaceutical Sciences*, 87(6):732–737, 1998.

Yannick Suter, Urspeter Knecht, Waldo Valenzuela, Michelle Notter, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. The LUMIERE dataset: Longitudinal Glioblastoma MRI with expert RANO evaluation. *Scientific Data*, 9(1):768, Dec 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01881-7. URL https://doi.org/10.1038/s41597-022-01881-7.

Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4ITK: Improved N3 Bias Correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010. ISSN 1558-254X. doi: 10.1109/tmi.2010.2046908. URL http://dx.doi.org/10.1109/TMI.2010.2046908.

Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL http://jmlr.org/papers/v9/vandermaaten08a.html.

Yibin Wang, William Duggar, David Caballero, Toms Vengaloor Thomas, Neha Adari, eswarakumar mundra, and Haifeng Wang. Brain Tumor Recurrence Prediction after Gamma Knife Radiotherapy from MRI and Related DICOM-RT: An Open Annotated Dataset and Baseline Algorithm (Brain-TR-GammaKnife), 2023a. URL https://www.cancerimagingarchive.net/collection/brain-tr-gammaknife/. Dataset. https://doi.org/10.7937/XB6D-PY67.

Yibin Wang, William Neil Duggar, David Michael Caballero, Toms Vengaloor Thomas, Neha Adari, Eswara Kumar Mundra, and Haifeng Wang. A brain MRI dataset and baseline evaluations for tumor recurrence prediction after Gamma Knife radiotherapy. *Scientific Data*, 10(1):785, Nov 2023b. ISSN 2052-4463. doi: 10.1038/s41597-023-02683-1. URL https://doi.org/10.1038/s41597-023-02683-1.

Sebastian F. Winter, Eugene J. Vaios, Alona Muzikansky, Maria Martinez-Lage, Marc R. Bussière, Helen A. Shih, Jay Loeffler, Philipp Karschnia, Franziska Loebel, Peter Vajkoczy, and Jorg Dietrich. Defining Treatment-Related Adverse Effects in Patients with Glioma: Distinctive Features of Pseudoprogression and Treatment-Induced Necrosis. *The Oncologist*, 25(8):e1221–e1232, June 2020. ISSN 1549-490X. doi: 10.1634/theoncologist.2020-0085. URL http://dx.doi.org/10.1634/theoncologist.2020-0085.

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, 2025a. URL https://arxiv.org/abs/2505.09965.

Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical World Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8319–8329, October 2025b.

Yuan Yin, Ibrahim Ayed, Emmanuel de Bézenac, Nicolas Baskiotis, and Patrick Gallinari. LEADS: Learning Dynamical Systems that Generalize Across Environments. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7561–7573. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/3df1d4b96d8976ff5986393e8767f5b2-Paper.pdf.

Jee Seok Yoon, Chenghao Zhang, Heung-Il Suk, Jia Guo, and Xiaoxiao Li. SADM: Sequence-Aware Diffusion Model for Longitudinal Medical Image Generation. In *Information Processing in Medical Imaging*, volume 13939 of *Lecture Notes in Computer Science*, pp. 388–400. Springer Nature Switzerland, 2023. ISBN 9783031340482. doi: 10.1007/978-3-031-34048-2_30. URL http://dx.doi.org/10.1007/978-3-031-34048-2_30.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, October 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Svetlana V. Zolotova, Andrey V. Golanov, Igor N. Pronin, Alexandra V. Dalechina, Anna A. Nikolaeva, Alexandra S. Belyashova, Dmitry Y. Usachev, Ekaterina A. Kondrateva, Polina V. Druzhinina, Boris N. Shirokikh, Talgat N. Saparov, Mikhail G. Belyaev, and Anvar I. Kurmukov. Burdenko's Glioblastoma Progression Dataset (Burdenko-GBM-Progression), 2023. URL https://www.cancerimagingarchive.net/collection/burdenko-gbm-progression/. Dataset. https://doi.org/10.7937/E1QP-D183.

SUPPLEMENTARY MATERIAL

Overview

- Section A: tumour-change detection, volume agreement, spatial patterns and qualitative examples.
- Section B: evaluation units, metric definitions, segmenter validation and statistical uncertainty.
- Section C: model architecture, training and inference, including the consistent-map derivation (Section C.9).
- Section D: dataset sources and access, cohort coverage and change statistics, eligibility, processing and quality control.
- Section E: protocols and statistics for the latent-space analysis in Figure 2.
- Section F: why the representation objective supervises change direction rather than the next state.
- Section G: extended related work on longitudinal representations, forecasting and treatment conditioning.

A ANALYSIS OF TUMOUR FORECASTS

A.1 CHANGE DETECTION AND VOLUME AGREEMENT

Figure S1 compares change detection, volume agreement and near-zero forecasts. Copy-input predicts no change for every case and ranks growth at chance; CounterSynth and DiffuseMorph also predict little change for many cases. Reporting these readouts together distinguishes detecting change from recovering its magnitude. LMWM ranks both endpoints best (growth AUC 76%, regression AUC 83%) and has the highest volume-change concordance. The ROC panels also distinguish growth from regression among the baselines: MambaControl ranks regression more accurately than growth (71% versus 65%), whereas Δ-LFM ranks growth better (71% versus 67%). Reporting both endpoints reveals differences that a single progression score would conceal. Evaluation units and statistical conventions are defined in Sections B.1 and B.2.

Volumes are measured by one fixed tumour segmenter applied to forecasts and acquired follow-up scans, using the voxel-count convention and reference threshold $V_{\min}$ defined in Appendix B.1. Growth is labelled when $\Delta V / \max(V_0, V_{\min}) \geq 25\%$ and $\Delta V \geq V_{\min}$, and regression when the corresponding quantities are at most -25% and $-V_{\min}$. Panel (b) marks the observed near-zero rate (6%) under $|\Delta V| < 2V_{\min}$. The 95% confidence intervals in panels (a,b) use 1,000 patient-bootstrap resamples. The figure and Table S1 score the same predictions; cohort metrics use case-bootstrap standard deviations in the table and patient-bootstrap intervals in the figure.

A.2 DISTRIBUTION OF PREDICTED CHANGE SIGNATURES

Figure 5 in the main paper compares the spatial patterns of observed and predicted change. The definitions below separate outcome alignment in signature space from coverage in the visualization.

Each change is represented by nine volumes: tissue lost inside the baseline tumour boundary in four distance bins (0–3, 3–6, 6–9 and $>$ 9 mm), and tissue gained outside it in five bins (0–3, 3–6, 6–9, 9–15 and $>$ 15 mm). Observed changes, all forecasting methods and the blur-null reference are embedded jointly with t-SNE (van der Maaten & Hinton, 2008) at perplexity 40.

For neighbourhood balanced accuracy (bAcc), each volume signature is transformed with asinh and divided by the standard deviation of the transformed observed signatures in each bin, plus 10^{-6}. Its five nearest observed signatures in this nine-dimensional space supply the outcome vote, using the growth and regression labels defined in Section A.1. The query case is excluded from the reference neighbours. Chance is 0.33 and observed signatures score 0.88. This measures neighbourhood alignment, distinct from the volume-threshold direction accuracy in Table S1. Gap is the mean Euclidean distance in the two-dimensional t-SNE embedding from each observed change to the method's nearest forecast. It measures how closely a method's forecasts cover the observed changes in the embedding.

Table S1: Segmentation-derived progression (%). LMWM has the highest follow-up Dice, growth AUC and volume-change concordance; copying the input retains high overlap while predicting no change.

		Overlap	Volume		Core change overlap		Direction / errors	
Method	Venue	Follow-up Dice ↑	Growth AUC ↑	ΔV CCC ↑	Growth Dice ↑	Regression Dice ↑	bAcc ↑	False growth ↓
Copy input	–	72.9±21.6	50.0±0.0	0.0±0.0	0.0±0.0	0.0±0.0	33.3±0.0	0±0
Blur input	–	69.5±20.4	47.8±4.2	−1.2±2.2	7.3±8.8	12.4±16.8	38.4±2.3	74±5
Δ-LFM	ICLR'26	66.1±21.5	71.5±3.5	39.5±5.5	14.5±14.2	27.5±24.0	53.4±2.8	76±6
MambaControl	arXiv'25	57.7±26.0	64.7±3.5	39.0±5.6	10.7±11.5	**30.6**±25.0	50.2±2.8	97±2
ImageFlowNet	ICASSP'25	71.1±21.3	47.5±4.0	4.2±1.6	3.9±8.3	13.7±20.3	34.1±2.2	17±5
GlioPREDICT	arXiv'26	57.7±24.7	59.6±4.1	16.2±4.4	8.8±10.8	24.8±20.4	42.8±3.1	89±4
CardiacAging	Front.CVM'22	64.3±23.7	43.6±4.5	8.6±3.1	3.7±6.5	17.5±19.3	36.7±2.8	42±6
CounterSynth	MedIA'23	72.3±21.4	46.1±4.0	−0.6±0.8	2.6±8.1	3.1±10.5	32.4±0.8	5±2
DiffuseMorph	ECCV'22	72.6±21.2	45.7±4.5	−0.4±0.8	2.3±8.4	2.6±12.2	33.5±1.1	**0**±0
MeWM	ICCV'25	64.8±24.2	50.6±4.3	4.1±2.4	6.0±7.8	13.9±17.1	33.8±2.8	68±6
BrLP	MICCAI'24	60.9±24.5	59.4±4.1	8.3±3.3	8.1±10.0	19.1±18.9	40.7±2.9	86±4
TADM	MICCAI'24	68.6±22.1	53.1±4.5	6.8±2.2	5.7±8.6	11.8±17.6	35.3±2.3	58±6
SADM	IPMI'23	59.3±24.6	53.7±7.9	12.0±7.7	6.2±7.7	26.3±19.0	40.7±5.0	85±8
TaDiff	TMI'25	70.0±21.7	56.9±7.6	6.1±5.2	11.2±11.2	17.3±15.7	38.3±4.9	70±10
LMWM°	Ours (Base)	70.4±21.7	64.9±3.8	23.3±5.3	9.8±14.4	17.1±20.2	50.6±2.8	38±6
LMWM†	Ours (+A)	72.9±23.2	68.8±3.2	40.7±6.4	13.9±12.6	25.1±15.8	53.5±3.1	48±7
LMWM	Ours (+A+H)	**75.7**±20.5	**76.2**±3.4	**47.6**±6.5	**18.3**±20.9	29.2±25.6	**55.0**±2.9	67±6

Follow-up Dice compares the fixed evaluator's masks on predicted and acquired follow-ups. Dice values are means ± case SD; core Dice includes cases with corresponding observed core change $\geq V_{\min}$. AUC, CCC, bAcc and false growth are estimates ± case-bootstrap SD. False growth is the fraction with predicted core growth $\geq V_{\min}$ among cases with observed core growth $< V_{\min}$. The reference threshold $V_{\min}$ is defined in Appendix B.1. Bold: best among learned methods. Base: current scan only; A: treatment field E; H: preceding latent change, consistent map and record text. Definitions: Appendix B.

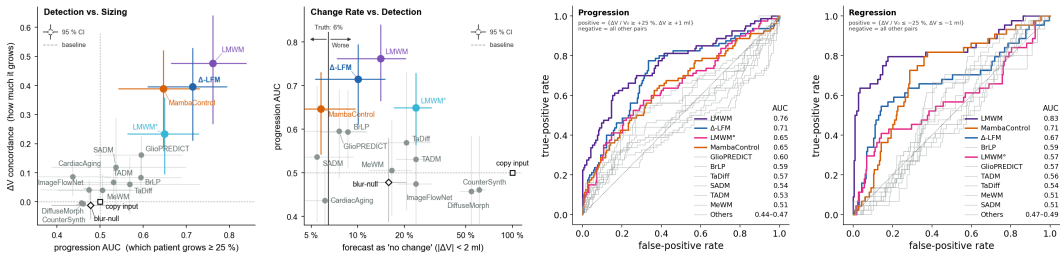

Figure S1: Ranking tumour change and recovering its magnitude require separate evaluation. **(a)** Growth AUC versus volume-change concordance. **(b)** Near-zero forecast rate versus growth AUC. **(c,d)** Growth and regression ROC curves. Crosses in (a,b): 95% patient-bootstrap confidence intervals.

A.3 Additional qualitative examples

Figure S2 complements Figure 6 with two further patterns of change. The upper example contains extensive central signal loss and a changed lesion boundary; several outputs preserve much of the baseline appearance despite substantial whole-lesion overlap. The lower example concentrates change near the lesion boundary, where an apparently similar image can still miss the location of signal loss. Together, the image, difference and contour views separate visual similarity from accurate reconstruction of change.

The panels use matched axial crops. Difference maps subtract the baseline image and share a case-specific scale: the 99th percentile of absolute observed change within the dilated baseline/follow-up tumour union, clipped to $[-1, 1]$ for display. The SSIM labels average a two-dimensional SSIM map

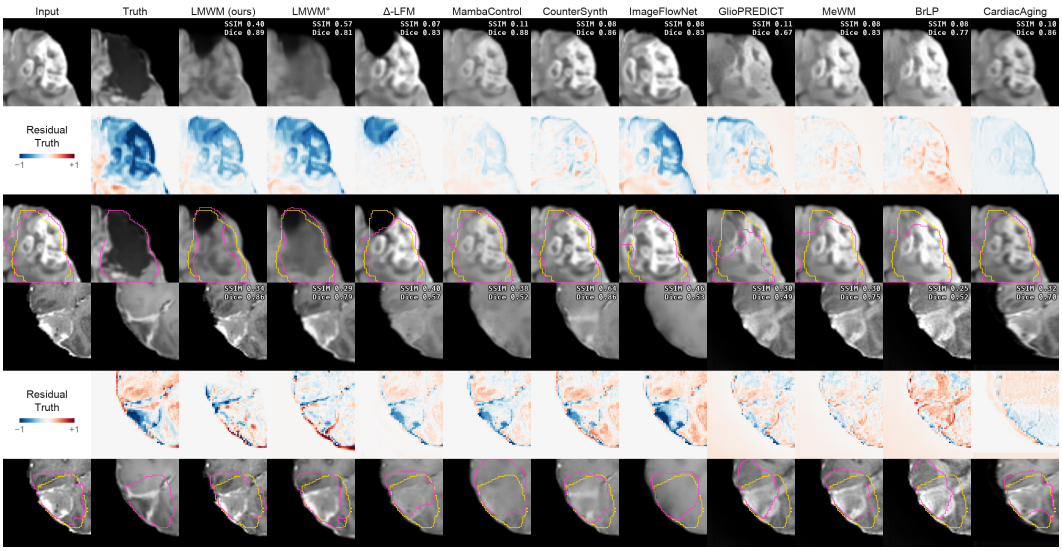

Figure S2: **Broad signal loss and boundary-localized change test different aspects of a forecast.** Rows: images, baseline-relative residuals (red: brighter; blue: darker) and tumour contours. Truth shows observed change. One fixed segmenter supplies both contours: yellow for the acquired follow-up and magenta for the displayed image (baseline in Input). Labels: tumour-region 2D SSIM and 3D Dice.

Table S2: Image fidelity and voxelwise change metrics. We introduce cPrec, cRec and cF1.

Metric	Definition	Optimum
PSNR	$10\log_{10}(r^2/\mathrm{MSE}_\Omega)$, with follow-up intensity range r	↑
SSIM	3D similarity, with voxels outside Ω zeroed	↑
LPIPS-3D	MedicalNet feature distance on a tumour-centred crop	↓
cF1	Harmonic mean of signed change precision P and recall R	↑
cPrec / cRec	$P = \mathrm{clip}_{[0,1]}(\langle d, \hat{d}\rangle_{\lvert\hat{d}\rvert>\tau}/\lVert\hat{d}\rVert^2_{\lvert\hat{d}\rvert>\tau})$; R swaps $d, \hat{d}$	↑
Spearman ρ	Rank correlation of median-centred change fields	↑

$d = x_1 - x_0$, $\hat{d} = \hat{x}_1 - x_0$; tumour region $\Omega = s_0 \cup s_1$. $\tau = \max\{0.25Q_{0.99}(|d|_\Omega), 10^{-6}\}$. Full definitions follow in this appendix.

within the tumour region of the displayed crop; Dice compares three-dimensional follow-up masks from the fixed evaluation segmenter.

B Evaluation setting and metrics

The GBM evaluation measures change localization, volume agreement and growth ranking through a fixed segmenter (Table S3), with complementary image-space diagnostics (Table S2; formulas below). These definitions support Tables 1 and S1.

B.1 Evaluation unit, image frame, and reference masks

An evaluation case is an adjacent scan pair for one modality. Write x_0, x_1, and $\hat{x}_1$ for its baseline, acquired follow-up, and predicted follow-up, and define

$$d = x_1 - x_0, \qquad \hat{d} = \hat{x}_1 - x_0, \qquad \Omega = \{s_0 > 0.5\} \cup \{s_1 > 0.5\}. \tag{8}$$

Here s_0, s_1 are the stored, brain-masked segmentations. Their union includes regions of both growth and regression. It is a scoring mask and never an inference input. All custom image metrics clip predictions to $[0, 1]$ before forming $\hat{d}$. Evaluation images are resampled from the $192 \times 192 \times 160$ grid at 1 mm to $1.5 \times 1.5 \times 1.25$ mm, padded or cropped to 128^3, and use a baseline-fitted intensity

window. For the segmentation-derived metrics below, volumes are expressed as voxel counts on this evaluation grid. The absolute change threshold is $V_{\min} = 8000/27 \approx 296.3$ voxels; a region meets this threshold when it contains at least 297 voxels.

T1 and contrast-enhanced T1 are scored as separate cases, and all methods are scored on the same test cases. Patient-level aggregation is defined in Section B.2.

B.2 Aggregation, uncertainty, and reproducibility

Per-case values are summarized by their arithmetic mean and sample standard deviation ($\mathrm{ddof} = 1$). Cross-case CCC, balanced accuracy, AUC and false-growth rate use point estimates plus the standard deviation of 300 case-resampling bootstrap replicates, with `numpy.random.default_rng(0)`.

Patient-level confidence intervals (Figures S1 and 2) resample patients with all their modalities and intervals, pairing methods on identical cases.

B.3 Evaluation segmenter and validity checks

The evaluation segmenter is a single-channel three-dimensional MONAI (Cardoso et al., 2022) U-Net with channel widths $(16, 32, 64, 128, 256)$, four stride-2 reductions, two residual units, instance normalization, and one output logit; logits above zero define tumour. It is trained on both visits of GBM training-split pairs with segmentations, with 8% of these entries held out for model selection; test patients are excluded from training and selection.

Training used binary cross-entropy plus soft Dice, AdamW with learning rate 3×10^{-4} and weight decay 10^{-5}, a 12-epoch cosine schedule, batch size 2, and mixed precision. Augmentations included Gaussian blur with $\sigma \sim U(0, 1.6)$ voxels, log-uniform gamma in $[0.6, 1.7]$, gain in $[0.6, 1.4]$, bias in $[-0.1, 0.1]$, noise standard deviation in $[0, 0.03]$, and independent axis flips with probability 0.5. The segmenter reaches a validation Dice of 66.6% and agrees with the stored follow-up annotations on the test cases at Dice $68.1 \pm 20.9\%$; all methods are scored with this same evaluator.

Reference predictions. Reference predictions are copy-input x_0, its 5^3 box-filtered image and the perfect target x_1.

The blur reference attains cF1 32.7% and Spearman 40.6%. Its core-growth and core-regression Dice are 7.3% and 12.4%, with a false-growth rate of 74%. The blur reference therefore sets the level against which segmentation-derived change scores are read. Translating the true follow-up change by six voxels still retains volume CCC 0.770, so we pair volume agreement with localization measures.

B.4 Geometry through a fixed image segmenter

Let S_{eval} be one frozen image-space segmentation model, distinct from the latent readout Q that the forecasting model uses internally. Define

$$A_0 = S_{\mathrm{eval}}(x_0), \qquad A_1 = S_{\mathrm{eval}}(x_1), \qquad \hat{A}_1 = S_{\mathrm{eval}}(\hat{x}_1), \qquad V_j = |A_j|. \tag{9}$$

One operator measures acquired and predicted images; stored annotations separately assess its accuracy.

Overlap and boundary-excluded changes. For nonempty $B \cup C$, $D(B, C) = 2|B \cap C|/(|B| + |C|)$. Follow-up Dice compares $\hat{A}_1$ with A_1. Growth regions are $G = A_1 \setminus A_0$ and $\hat{G} = \hat{A}_1 \setminus A_0$; regression regions are $R = A_0 \setminus A_1$ and $\hat{R} = A_0 \setminus \hat{A}_1$. Their Dice scores use these respective pairs. When the true change region is empty, the rule returns 1 if the predicted region is smaller than $V_{\min}$, and 0 otherwise; two empty regions therefore score 1.

Let $\mathcal{D}_2$ and $\mathcal{E}_2$ denote two iterations of dilation and erosion with a six-connected three-dimensional structuring element. Define

$$\begin{aligned} G_{\mathrm{core}} &= A_1 \setminus \mathcal{D}_2(A_0), & \hat{G}_{\mathrm{core}} &= \hat{A}_1 \setminus \mathcal{D}_2(A_0), \\ R_{\mathrm{core}} &= \mathcal{E}_2(A_0) \setminus A_1, & \hat{R}_{\mathrm{core}} &= \mathcal{E}_2(A_0) \setminus \hat{A}_1. \end{aligned} \tag{10}$$

Table S3: Spatial change, volume agreement and growth direction measure distinct aspects of tumour progression. All measures use segmented tumour masks.

Metric	Target	Definition or interpretation
Follow-up Dice	↑	$D(\hat{A}_1, A_1)$
Core growth / regression Dice	↑	Overlap beyond the specified baseline boundary band
ΔV CCC	↑	Cross-case concordance of volume change
Direction bAcc	↑	Mean recall over growth, stable, and regression classes
False-growth rate	↓	Predicted core growth among cases without true core growth
Volume-growth / regression AUC	↑	Ranking under the 25% and $V_{\min}$ volume criteria (Appendix B.1)

The boundary band consists of two six-connected voxel steps. Core Dice is calculated only when the corresponding true core-change volume is at least $V_{\min}$. Other cases are excluded from that mean. False-growth rate is

$$\mathrm{FGR} = \frac{\sum_{i:\,|G_{\mathrm{core},i}|<V_{\min}} \mathbf{1}\{|\hat{G}_{\mathrm{core},i}| \geq V_{\min}\}}{|\{i : |G_{\mathrm{core},i}| < V_{\min}\}|}. \tag{11}$$

Volume agreement and direction. For case i, let $t_i = |A_{1,i}| - |A_{0,i}|$ and $p_i = |\hat{A}_{1,i}| - |A_{0,i}|$. Across valid cases,

$$\mathrm{CCC} = \frac{2\,\mathrm{cov}(t,p)}{\mathrm{var}(t) + \mathrm{var}(p) + (\bar{t} - \bar{p})^2}. \tag{12}$$

CCC uses population moments. Direction labels apply the same case-specific threshold $h_i = 0.1\max(V_{0,i}, V_{\min})$ to t_i and p_i: growth above h_i, regression below $-h_i$, and stable otherwise. Direction balanced accuracy is the mean recall over these three classes. An all-stable predictor scores $1/3$ when all three true classes occur.

B.5 Volume-based ranking across cases

Define $r_i = t_i/\max(V_{0,i}, V_{\min})$ and $\hat{r}_i = p_i/\max(V_{0,i}, V_{\min})$. The growth AUC (progression AUC in Figure S1) uses the binary label $r_i \geq 0.25$ and $t_i \geq V_{\min}$, with score $\hat{r}_i$. Regression AUC uses $r_i \leq -0.25$ and $t_i \leq -V_{\min}$, with score $-\hat{r}_i$. These labels are segmentation-derived volume categories.

For either binary endpoint, let R_i be the average rank of its prediction score, with n_1 positive and n_0 negative cases. Then

$$\mathrm{AUC} = \frac{\sum_{i:y_i=1} R_i - n_1(n_1+1)/2}{n_1 n_0}. \tag{13}$$

It is undefined if either class is absent.

B.6 Image appearance and change fields

PSNR. Let $r = \max(x_1) - \min(x_1)$ be the acquired follow-up range over the full volume and $\mathrm{MSE}_\Omega = |\Omega|^{-1}\sum_{v\in\Omega}(\hat{x}_{1,v} - x_{1,v})^2$. Then

$$\mathrm{PSNR}_\Omega = 10\log_{10}\frac{r^2}{\mathrm{MSE}_\Omega}. \tag{14}$$

SSIM and perceptual distance. Tumour SSIM is the mean output of MONAI's three-dimensional `SSIMMetric`, with data range 1, after setting both images to zero outside Ω. Shared zero background contributes to this volume-averaged score. Three-dimensional tumour SSIM saturates at 0.985–0.990 for every method, so the result tables report PSNR and LPIPS-3D instead; the qualitative figures label each panel with a two-dimensional SSIM averaged within the tumour region of the displayed crop.

The reported LPIPS-3D uses MONAI's three-dimensional perceptual loss, `is_fake_3d=False`, and the MedicalNet ResNet-10 (Chen et al., 2019) pretrained on 23 datasets. It uses the implementation's MedicalNet intensity normalization. The tumour variant extracts the same 64^3 crop from prediction and target, centred at the integer-rounded centroid of Ω and shifted to remain inside the volume. It is a lower-is-better feature distance; reported values are multiplied by 10^3. Copy-input references use x_0 in place of $\hat{x}_1$ through the identical computation.

Signed change F1. We define cPrec and cRec as clipped signed projections and combine them using the standard F1 harmonic mean. This signed-change formulation is introduced in this work. Let $C(a) = \min(1, \max(0, a))$. Within Ω, the canonical active sets use a common target-derived threshold:

$$\tau = \max\{0.25Q_{0.99}(|d|_\Omega), 10^{-6}\}, \quad R = \{v \in \Omega : |d_v| > \tau\}, \quad P = \{v \in \Omega : |\hat{d}_v| > \tau\}. \tag{15}$$

The signed projection scores and their harmonic mean are

$$r_c = C\left(\frac{\sum_{v\in R} d_v \hat{d}_v}{\sum_{v\in R} d_v^2 + \epsilon}\right), \qquad p_c = C\left(\frac{\sum_{v\in P} d_v \hat{d}_v}{\sum_{v\in P} \hat{d}_v^2 + \epsilon}\right),$$
$$\mathrm{cF1} = \frac{2p_c r_c}{p_c + r_c + \epsilon}, \qquad \epsilon = 10^{-9}. \tag{16}$$

cPrec and cRec denote p_c and r_c for the raw fields. Empty active sets contribute zero. The projection scores penalize sign reversal; voxelwise Spearman ρ complements them with rank-based agreement. Ignoring stabilization, a non-degenerate perfect forecast scores 1, copying and sign reversal score 0, and $\hat{d} = 0.5d$ or $2d$ scores $2/3$ when the active sets are nonempty.

Rank correlation. Subtract the median within Ω separately from each field, obtaining $\tilde{d}$ and $\tilde{\hat{d}}$. Voxelwise Spearman ρ is the Pearson correlation of their average ranks. A predicted field with peak-to-peak range below 10^{-8} receives zero Spearman correlation.

C Implementation Details

This section describes the implemented LMWM in Figure 3. LMWM denotes the full model with both treatment and history modules; LMWM$^\circ$ and LMWM† are its nested variants (Section 4). The parts are fitted in order: representation, shared transition, treatment modules, history modules. The clinical-text encoder is fixed.

C.1 Autoencoder Architecture and Optimization

The representation uses a single-channel three-dimensional MAISI autoencoder (Guo et al., 2025); T1 and T1ce are separate samples for the same encoder. Four latent channels represent each image, with 112^3 training patches mapping to 4×28^3 codes and 128^3 inference inputs mapping to 4×32^3 codes. We use posterior means for longitudinal differences.

The autoencoder starts from the public MAISI weights and is first fine-tuned with the reconstruction, perceptual, adversarial and KL terms of Eq. 17 at learning rate 10^{-4} for 12,000 steps; change-direction training then continues from this autoencoder and its discriminator. AdamW uses learning rate 10^{-5} and weight decay 10^{-6} for the autoencoder, with a tenfold learning rate for the affine change predictor. Batch size is one longitudinal tuple; training uses bfloat16 and activation checkpointing. Online weight averaging uses decay 0.999, and the final representation averages five epoch-level averaged checkpoints uniformly. Representation training uses no segmentation head or segmentation loss.

Training resamples the canonical images to $1.5 \times 1.5 \times 1.25$ mm and extracts tumour-centred 112^3 crops. Intensities are mapped to $[0, 1]$, then standardized per volume for encoding; image losses use the restored $[0, 1]$ scale. All visits in a tuple share the crop and left–right flip (probability 0.5); rigid augmentation is disabled. A shared background outside a dilated, softened foreground union isolates lesion change during training. Two-visit patients repeat their last visit for reconstruction and are excluded from prediction supervision.

C.2 INHERITED REPRESENTATION OBJECTIVES

Representation training uses

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{rec}} + 0.3\mathcal{L}_{\text{perc}} + 10^{-7}\mathcal{L}_{\text{KL}} + 0.1\eta(n)\mathcal{L}_{\text{adv}} + 0.3\mathcal{L}_{\text{chgrec}}. \tag{17}$$

Reconstruction uses the Charbonnier penalty with $\epsilon = 10^{-3}$. Perceptual supervision uses the SqueezeNet-based loss (Zhang et al., 2018); adversarial supervision uses a least-squares patch discriminator (Mao et al., 2017) with ramp $\eta(n) = \min\{1, (n-1)/100\}$ over steps n. No segmentation or SSIM loss is used.

For reconstructed and observed first-to-last changes $\delta\tilde{x}$ and δx, the image-change term is

$$\mathcal{L}_{\text{chgrec}} = \frac{\sum_x A(x)|\delta\tilde{x}(x) - \delta x(x)|}{\max\{\sum_x A(x)|\delta x(x)|, 10^{-6}\}}, \tag{18}$$

averaged over tuples. The high-change mask A thresholds the 90th percentile of absolute change after non-brain entries receive a negative sentinel, then intersects with the brain.

C.3 CHANGE-DIRECTION OBJECTIVE

The encoder Enc_ψ and decoder D_ω map between images and disease-state representations. We use the posterior mean $z_i = \text{Enc}_\psi(x_i)$ so that temporal differences exclude encoder sampling noise. Its four channels are learned coordinates, without a prescribed correspondence to tissue labels. The inherited objective $\mathcal{L}_{\text{base}}$ preserves image appearance and observed changes (Eq. 17). We add prediction of future latent-change direction.

For three distinct consecutive visits $(i-1, i, i+1)$ within one patient and modality, ordered chronologically, an auxiliary predictor receives the current state and preceding change and estimates $d_i = z_{i+1} - z_i$ from $[z_i,\ z_i - z_{i-1}]$ (Section 2.2). The predictor is a spatially local affine map: a $1 \times 1 \times 1$ convolution from eight channels to four, with a bias. It learns under the observed interval distribution without elapsed-time or treatment inputs; those condition the later dynamics model. Reliably ordered visits can therefore supervise the representation even when exact dates are unavailable. Let M_i be the union of current and target foreground segmentations, mapped to a binary latent mask by adaptive max pooling and used only in training. The mask weights the loss and is excluded from the predictor inputs. For a spatial mask W, define the channel-averaged directional agreement

$$C_W(u, v) = \frac{1}{4}\sum_{k=1}^{4} \frac{\sum_x W(x)u_k(x)v_k(x)}{N_W(u_k, v_k)}. \tag{19}$$

Here N_W is the product of masked spatial norms, with the numerical floors given below. Averaging normalized channel agreements gives each channel equal weight in the loss. Let $\bar{d}$ be a detached exponential moving average of training change fields, initialized to zero, and define $\ell_W(i) = 1 - C_W(\hat{d}_i, d_i) + C_W(\bar{d}, d_i)$. Minimizing the first term aligns the predicted and observed changes. The signed mean-direction term discourages positive alignment of target changes with the running mean and can reward opposite alignment. If $\mathcal{V}$ indexes valid triplets and $\mathcal{V}_M = \{i \in \mathcal{V} : \sum_x M_i(x) \geq 8\}$, the whole-field and tumour losses are averaged over their respective eligible samples:

$$\mathcal{L}_{\text{pred}} = \frac{\sum_{i\in\mathcal{V}} \ell_1(i)}{\max(1, |\mathcal{V}|)} + \frac{\sum_{i\in\mathcal{V}_M} \ell_{M_i}(i)}{\max(1, |\mathcal{V}_M|)}, \tag{20}$$

The ramp $\chi(n)$, weight λ_{pred} and moving-average update are specified below. Gradients reach both predicted and target codes, so the mean-direction term shapes the encoder despite the detached running mean. After representation learning, the auxiliary predictor is discarded and the encoder and decoder are frozen; all fitting respects the patient exclusions in Appendix C.11.

The distinction lies in directional supervision. An affine next-state predictor has the same expressivity as the affine change predictor g (Appendix F) after adding z_i, and their squared prediction errors coincide under this reparameterization. Cosine agreement changes under this shift, but is insensitive to positive rescaling away from the numerical floors. Accurate direction therefore leaves change magnitude to be tested independently in decoded forecasts. Appendix F makes this distinction precise.

C.4 PREDICTION LOSS AND TRAINING TUPLES

Here $\chi(n) = \min(1, n/500)$ for one-based step n, and $\lambda_{\text{pred}} = 0.1$. The tumour term has unit relative weight; masks smaller than eight latent voxels contribute only to the whole-field term. After computing the loss, the moving average is updated with decay 0.99 using the detached mean change over $\mathcal{V}$; an empty valid set leaves it unchanged. Target codes retain encoder gradients. The implemented norm floors in Eq. 19 are as follows. With $S_W(u) = \sum_x W(x)u(x)^2$, the whole-field term uses $N_1(u, v) = \max\{\sqrt{S_1(u)}, 10^{-8}\} \max\{\sqrt{S_1(v)}, 10^{-8}\}$. The tumour term uses

$$N_M(u, v) = \max\Big\{\sqrt{\max\{S_M(u), 10^{-12}\}}\sqrt{\max\{S_M(v), 10^{-12}\}}, 10^{-8}\Big\}. \tag{21}$$

Consecutive triplets share patient, modality, spatial crops and geometric augmentations. Two-visit patients supply (x_0, x_1, x_1) for reconstruction; the duplicate endpoint excludes prediction supervision. Reliable order suffices for the time-free auxiliary predictor; dynamics fitting requires verified intervals. The autoencoder blends visits with a shared background outside a softened union of training segmentations. Evaluation uses acquired images without this training-time background blending.

C.5 FORECAST INPUTS AND TREATMENT RECORDS

Information regimes. Forecasts condition on the recorded interval course. Historical residuals and query forecasts use the same information. Target images, segmentations and endpoint-derived spatial supports serve only as supervision.

Treatment-record provenance. Each course retains class, dated events, available dose or drug information, source maps and availability flags. Recorded dates and bounded timing remain distinct. Pre-baseline courses contribute their remaining response to the query interval. Without an eligible dated event, the corresponding calendar and treatment field are zero.

C.6 SHARED TRANSITION AND CONTEXT

A 2^3 latent patch produces each of 4,096 tokens. A two-layer patch embedding uses a 128-dimensional bottleneck and model width 768. Twelve blocks each contain twelve-head self-attention, cross-attention and a SwiGLU layer with hidden width 2,048. RMS normalization, query–key normalization and 3D rotary positions support the self-attention path. Cross-attention uses learned positions in the context memory. A linear head reconstructs four latent channels.

The context memory consists of 512 segmentation tokens and 512 preceding-change tokens, each of width 256. Convolutional encoders reduce the input segmentation and latent change to 8^3 spatial grids; a learned missing-history token replaces the preceding change in LMWM$^\circ$ and LMWM†, and when unavailable in the full model. Their spatial positions are learned. The broadcast condition combines flow-time, horizon, age and sex embeddings and reaches every block's adaptive normalization and the output normalization. The horizon transform is $\text{clip}_{[0,1]}(\log(1+\Delta t)/\log(1+1095))$ with Δt in days.

The transition has 169,857,568 backbone parameters, 9,919,552 context-encoder parameters and 606,720 scalar-embedding parameters. Its context memory also carries the three scalars as tokens, giving 1,027 tokens; Figure 3 shows the 1,024 image-derived tokens.

C.7 TREATMENT FIELD AND BLOCKWISE ADAPTATION

Spatial and plan inputs. For class $k \in \{S, R, C\}$, N_k encodes current anatomy, tissue segmentation, a source treatment map and plan attributes p_k into eight channels on the 16^3 token grid. Surgery uses resection support and distance to that support; radiotherapy uses dose and fractionation-derived information; systemic therapy uses brain support and drug attributes. Plan attributes carry missingness indicators. Spatial encoding receives no absolute dates. Treatment timing is represented by the calendar below.

For systemic therapy, p_C concatenates a six-category one-hot code (concurrent temozolomide, adjuvant temozolomide, bevacizumab, tumour-treating fields, other and unknown), the fraction of inter-

val days on treatment, a schedule-unknown flag and a missingness flag. The categories distinguish drug and regimen; tumour-treating fields is retained as a separate adjunctive treatment. These attributes are broadcast over the image grid and concatenated with anatomy and brain support before N_C produces the spatial response code. This code represents learned treatment-response features.

Response functions. The peak-normalized kernels are

$$\kappa_S(u;T) = e^{-u/T}, \quad \kappa_R(u;T) = \frac{u}{T}e^{1-u/T}, \quad \kappa_C(u;T) = \frac{(u/T)^2}{4}e^{2-u/T}, \tag{22}$$

for $u \geq 0$, and zero otherwise. These forms share the gamma-delay structure of a serial transit model (Krzyzanski, 2011). With m identical first-order stages, initial state $q_1(0) = 1$, $q_{j>1}(0) = 0$, and

$$\dot{q}_1 = -q_1/T, \qquad \dot{q}_j = (q_{j-1} - q_j)/T \quad (j = 2, \ldots, m), \tag{23}$$

the final stage is $q_m(u) = e^{-u/T}(u/T)^{m-1}/(m-1)!$. Dividing by its peak gives Eq. 22 for $m = 1, 2, 3$, respectively. Surgery therefore has an immediate peak, whereas the radiotherapy and systemic-treatment kernels peak at T and $2T$. The stage count controls the onset shape; the learned time scale controls its duration. This is the gamma family with mean delay mT and mode $(m-1)T$ for $m > 1$. For each token and channel, the response function is

$$h_k(u;\theta,p) = \kappa_k(u;T_k(\theta,p)) \odot [1 + f_k(\log(1+u),\theta,p)]\,. \tag{24}$$

The correction f_k is a two-layer MLP with hidden width 64, tanh activation and zero-initialized final layer. The bounded time-constant parameterization is $T_k = T_{k,\text{init}} \odot \exp(\text{clip}_{[-2,2]}(W_{T,k}[\theta,p]))$, with $W_{T,k} = 0$ initially. Postoperative cavities in brain metastases can change over the weeks preceding radiosurgery (Shah et al., 2016a), while radiation-associated MRI changes in glioma include early pseudoprogression and later treatment-induced necrosis (Winter et al., 2020). We represent these different time scales within the same kernel family. Initial times are 21 and 120 days across the two four-channel surgery groups, 30 and 180 days for radiotherapy, and 10 days for all chemotherapy channels. These values initialize the learned responses; the cited studies motivate their temporal structure.

Calendars. The fixed grid $u_g = \exp(g\log(1501)/63) - 1$, $g = 0, \ldots, 63$, spans 0–1500 days. Dated surgery, radiotherapy fractions and systemic-treatment days contribute impulses. For each day of the forecast interval, time since each eligible impulse is assigned by linear interpolation on $\log(1+u)$; averaging across interval days gives w_{ikg}. The weights are normalized once against a reference course evaluated with the initial kernel. They have no spatial index or learned parameters. Equation 3 combines them with the spatial response functions.

Two treatment operations. Let $v_{\ell j}$ denote residual-stream features entering sub-block j of block ℓ, and let $g_{\ell j}, b_{\ell j}, a_{\ell j}$ be the fixed backbone's adaptive scale, shift and residual gate. The projector E_A, pooling followed by an MLP, first maps the field to treatment tokens; E below denotes these tokens. Three shared dictionaries $D_g, D_b, D_w \in \mathbb{R}^{768\times 8}$ map the field to

$$g^E_{\ell j} = D_g(E \odot \gamma_{\ell j}), \qquad b^E_{\ell j} = D_b(E \odot \beta_{\ell j}), \tag{25}$$

$$y_{\ell j} = \text{RMSNorm}(v_{\ell j}) \odot (1 + g_{\ell j} + g^E_{\ell j}) + b_{\ell j} + b^E_{\ell j}, \tag{26}$$

$$v_{\ell,j+1} = v_{\ell j} + a_{\ell j} \odot \mathcal{B}_{\ell j}(y_{\ell j}), \qquad v_{\ell+1,1} = v_{\ell,4} + D_w(E \odot \omega_\ell). \tag{27}$$

The three $\mathcal{B}_{\ell j}$ are self-attention, cross-attention and the feed-forward sub-block. Gains $\gamma_{\ell j}, \beta_{\ell j} \in \mathbb{R}^8$ are learned per sub-block and $\omega_\ell \in \mathbb{R}^8$ per block, and the dictionaries are bias-free, so $E = 0$ leaves every block unchanged. Modulation enters before each attention calculation and passes through its existing residual gate. The block-output residual is added after all three sub-blocks and bypasses those gates. The field enters additively at these projections.

The treatment module has 12,864 modulation parameters and 6,240 block-output parameters, totaling 19,104, excluding E_A and the spatial and response encoders. Shared dictionaries constrain each operation to eight feature directions.

C.8 HISTORY ENCODING AND PARALLEL ATTENTION

A fixed Bio_ClinicalBERT encoder (Alsentzer et al., 2019) reads the record's text fields; a projector E_H, pooling followed by an MLP, maps the consistent map to tokens. Their tokens form history memory h, with spatial positions retained for map tokens. Each block learns independent $W^H_{k,\ell}, W^H_{v,\ell}$ and scalar λ_ℓ, as in Eq. 6. The combined attention is passed through the original output projection and the backbone's existing cross-attention residual gate. Original context keys and values remain unchanged.

Text embeddings have width 768, whereas the original context memory has width 256. Direct 768-to-768 key and value projections require 14,155,776 weights across twelve blocks, plus twelve gates; using width 256 requires 4,718,592 projection weights. These counts exclude biases and E_H. History tokens are projected to the context width 256 before the key–value projections. The record is serialized as one sequence per patient, one clause per event in chronological order, giving the class, the day relative to the forecast origin (or *undated*) and the available extent, dose or drug; all output tokens are kept.

Each gate λ_ℓ is initialized to zero. Complete absence of map and text bypasses the branch and recovers the treated transition; a zero map with nonempty text still supplies history tokens. With the treatment field also zero and no earlier scan, the model reduces to the treatment-agnostic transition LMWM° (Section 4).

C.9 CONSISTENT MAP: CONSTRUCTION AND RATIONALE

The consistent map summarizes the component of a patient's past evolution that repeatedly differs from the treatment-conditioned forecast. Its inputs are completed, non-overlapping intervals before the query origin. All residuals use the same encoder, treated transition and registered spatial grid.

Residual rates. For interval j, subtract the treated endpoint prediction from the acquired endpoint latent and divide by elapsed time:

$$e_j(x) = z_{j+1}(x) - F_E(z_j; \phi_j, E_j)(x), \qquad r_j(x) = \frac{e_j(x)}{\Delta_j \, \sigma_r(x)}. \tag{28}$$

Here $\sigma_r(x) > 0$ is a fixed scale in latent units per day, shared across channels. Normalization makes the map dimensionless and keeps high-variance locations from dominating its encoding. We use the voxelwise standard deviation of residual rates on held-out fitting intervals, shared across channels and floored at 10^{-6}. The scale is fixed before history adaptation. Using a treated prediction makes r_j a deviation from the model's expected response to that interval's recorded care. Dividing by Δ_j expresses deviations from different observation intervals on a common rate scale.

Duration weighting and short-history shrinkage. At one location x, write $v(x)$ for a persistent, normalized residual rate. Consider the working noise model

$$\frac{e_j(x)}{\sigma_r(x)} = \Delta_j v(x) + \eta_j(x), \qquad \eta_j(x) \sim \mathcal{N}(0, \sigma^2 \Delta_j I), \quad v(x) \sim \mathcal{N}(0, \sigma^2 T_0^{-1} I). \tag{29}$$

It assumes residual-increment variance proportional to duration and independent interval noise conditional on v. Thus r_j has precision proportional to Δ_j. The posterior mean and mode coincide and minimize

$$\frac{1}{2} \sum_{j<i} \Delta_j \|r_j - v\|_2^2 + \frac{T_0}{2} \|v\|_2^2. \tag{30}$$

Setting the gradient to zero gives

$$\tilde{v}_i = \frac{b_i}{T_i + T_0} = \frac{T_i}{T_i + T_0} \bar{r}_i, \qquad b_i = \sum_{j<i} \Delta_j r_j, \quad T_i = \sum_{j<i} \Delta_j, \quad \bar{r}_i = \frac{b_i}{T_i}. \tag{31}$$

The first factor measures accumulated observation duration relative to the prior strength. With $T_0 =$ 90 days, 90 days of history gives a factor of one half; the factor approaches one as history grows. Empty history gives zero directly. This working model for forecast residuals motivates duration weighting and shrinkage. All durations are in days.

Directional coherence. Shrinkage controls how much history is available. A second factor measures whether its residual vectors agree:

$$a_i(x) = \sum_{j<i} \Delta_j \|r_j(x)\|_2, \qquad c_i(x) = \frac{\|b_i(x)\|_2}{a_i(x) + \epsilon}, \qquad H_i(x) = c_i(x)\tilde{v}_i(x). \tag{32}$$

The triangle inequality gives $0 \le c_i \le 1$. Ignoring the numerical stabilizer when $a_i > 0$, aligned vectors have $c_i = 1$; equal opposing contributions give $c_i = 0$. More generally, writing $w_j = \Delta_j \|r_j\|_2$ and $u_j = r_j / \|r_j\|_2$ for nonzero residuals,

$$\|b_i\|_2^2 = \sum_j w_j^2 + 2 \sum_{j<k} w_j w_k \, u_j^\top u_k. \tag{33}$$

Coherence therefore summarizes agreement between residual directions, weighted by duration and magnitude. It adds attenuation when disagreement already reduces the vector mean. For example, two equal-duration residuals v and $-v$ cancel; two copies of v preserve the direction and receive only the short-history shrinkage. The coherence gate is applied to the posterior mean of Eq. 31.

Computation and use. For each eligible interval, accumulate b_i, a_i and T_i once on the fixed patient grid; then evaluate Eq. 32. Channel norms are voxelwise. The positive stabilizer ϵ has units of days under the dimensionless rate convention and gives $H_i = 0$ when every residual is zero. Overlapping all-forward pairs are excluded to avoid counting observations repeatedly. The map is updated when an acquired scan completes a new interval. Its spatial encoder and the record-text encoder provide the history tokens of Section 2.5.

With a single interval, coherence is near one and the duration shrinkage alone sets the map's magnitude.

C.10 Transition and adaptation training

Transition objective. The transition is trained by flow matching on the straight path $z^t = (1 - t)\, z_i + t\, z_{i+1}$ with $t \sim \mathcal{U}[0, 1]$, whose velocity is the residual $z_{i+1} - z_i$:

$$\mathcal{L}_{\text{flow}} = \frac{1}{4|\mathcal{G}|} \sum_{x \in \mathcal{G}} \big\| v_\theta(z^t, t; \phi_i)(x) - \big(z_{i+1}(x) - z_i(x)\big) \big\|_1, \tag{34}$$

unweighted over the latent grid $\mathcal{G}$, plus a least-squares latent adversarial term of weight 0.1 and the Dice loss of the lesion readout Q against the follow-up segmentation. Inference integrates v_θ from $t = 0$ to $t = 1$ with five Euler steps.

AdamW uses learning rate and weight decay 10^{-4}, batch size two, 1,000-step warm-up, gradient clipping at one, mixed precision and seed 42. Weight averaging uses decay 0.999. The latent discriminator uses two levels with initial width 32 and eight input channels, AdamW at 10^{-4} with $\beta = (0.5, 0.9)$, and starts after 1,000 generator updates.

Adaptation stages. Treatment adaptation uses the same loss on the frozen transition, with AdamW at 10^{-4} for dictionaries and gains and 5×10^{-5} for spatial encoders and response functions, with batch size two, bfloat16 and weight averaging at 0.999. Treatment classes alternate across steps. History training holds the treated transition fixed and fits only the map encoder, history projections and gates on queries following completed support intervals.

$$\hat{z}_{i+1}\big|_{\lambda=0} = F_E(\,\cdot\,; E), \qquad \hat{z}_{i+1}\big|_{\lambda=0, E=0} = F_\theta(\,\cdot\,). \tag{35}$$

Without an earlier scan, the preceding-change context takes its missing-history token, and Eq. 35 yields LMWM° exactly.

C.11 Patient splits and residual fitting

The forecasting experiments use GBM patients. Before quality control and task-specific exclusions, the patient-level split assigns 652 patients to training, 82 to validation and 82 to testing (approximately 80/10/10; seed 42), grouping all visits, modalities and history windows from each person.

Table S4: Dataset sources, official links and license/access terms for the eight glioma and four metastasis collections.

Dataset	Source	Link	License / access
Glioma			
UCSF-ALPTDG (Fields et al., 2024)	UCSF	DOI	Non-commercial use under a data use agreement (DUA).
Burdenko (Zolotova et al., 2023)	TCIA	DOI	NIH controlled access (images); CC BY 4.0 (clinical data).
MU-Glioma-Post (Mahmoud et al., 2025; Dhemesh et al., 2025)	TCIA	DOI	CC BY 4.0.
LUMIERE (Suter et al., 2022)	Figshare	DOI	CC0 (MRI and segmentation release).
RHUH-GBM (Cepeda et al., 2023a;b)	TCIA	DOI	CC BY 4.0 (brain-extracted NIfTI and clinical data); NIH controlled access (DICOM).
Ivy GAP (IVYGAP2) (Shah et al., 2016b)	TCIA	DOI	NIH controlled access (images).
UCSD-PTGBM (Gagnon et al., 2026; 2025)	TCIA	DOI	CC BY 4.0.
SAILOR (Hovden et al., 2023)	EBRAINS	DOI	EBRAINS DUA; dataset attribution required.
Brain metastasis			
Yale (Chadha et al., 2025b;a)	TCIA	DOI	CC BY 4.0.
PROTEAS (Flouri et al., 2025)	Zenodo (v3)	DOI	CC BY 4.0.
RFUDS (Fernandes et al., 2024)	Zenodo	DOI	CC BY 4.0.
Brain-TR-GammaKnife (Wang et al., 2023b;a)	TCIA	DOI	NIH controlled access (images); CC BY 4.0 (clinical data).

TCIA: The Cancer Imaging Archive. Terms as of September 2026. Where image and clinical releases carry different terms, redistribution of derived data follows the applicable source agreement.

Five-fold patient-level cross-validation within the training set selects adaptation settings. Residual scales use held-out fitting intervals and never test patients. Residual generation uses a transition that did not fit the corresponding patient's interval: within the training set, fold-held-out patients supply the residuals from which $\sigma_r(x)$ and the history branch are fitted. Inference uses one model; no fold averaging is used.

C.12 Inference grid and image readout

The inference tensor is 128^3 and the latent grid is 32^3. Images are resampled to $1.5 \times 1.5 \times 1.25$ mm and padded or cropped to 128^3, giving latent spacing $6 \times 6 \times 5$ mm and token spacing $12 \times 12 \times 10$ mm. Images used in longitudinal residuals share the patient reference frame and intensity convention.

The readout Q supplies $\hat{s}_{i+1} = Q(\hat{z}_{i+1})$. Equation 7 uses this predicted mask directly, without dilation or union with the current lesion. The decoder is evaluated at both latents, their difference is masked, and the current image is added back. Q is a latent segmentation head trained with the transition by a Dice loss against the follow-up segmentation; its output is upsampled to the image grid.

D Data curation and quality control

D.1 Dataset sources and access

Table S4 lists the eight glioma and four brain metastasis collections, with their source references, official links and license/access terms.

D.2 Populations and eligibility

The resource covers glioma of different grades and brain metastases. All images originate from existing de-identified research collections and are harmonized with explicit task eligibility.

Table S5: Cohort coverage, longitudinal depth and treatment records across twelve collections. Earlier scans are available for 59% of glioma intervals and 83% of metastasis intervals.

				Δt, days	History	4 seq.	4 acq.	GT seg.	Treatment in interval			
Collection	Records (test)	Visits	Pairs (test)	median [IQR]	(%)	(%)	(%)	(%)	S	R	C	U
Glioma												
UCSF	297 (30)	590	293 (30)	64 [56, 106]	0	100	100	0	7	1	257	–
MU	143 (15)	456	306 (26)	69 [45, 93]	58	0	0	100	–	22	220	–
Burdenko	141 (15)	332	180 (23)	69 [52, 80]	46	77	68	0	–	76	–	–
LUMIERE	85 (9)	554	466 (52)	77 [28, 98]	82	95	0	0	100	29	11	–
RHUH	40 (4)	71	31 (2)	7 [3, 10]	0	100	100	100	31	–	31	–
IVYGAP2[b]	34 (3)	263	202 (10)	56 [33, 70]	86	97	92	10	25	–	–	–
UCSD	33 (3)	77	44 (5)	113 [58, 190]	25	100	100	100	–	–	17	–
SAILOR	27 (3)	270	243 (16)	76 [14, 91]	89	99	99	100	–	102	213	41
GBM total	800 (82)	2,613	1,765 (164)	68 [39, 92]	59	78	56	34	163	230	749	41
Brain metastases												
Yale	906 (93)	6,448	5,451 (451)	84 [54, 124]	84	100	56	0	–	–	–	582
PROTEAS	45 (4)	178	125 (11)	87 [49, 107]	66	100	100	0	–	42	18	–
RFUDS	39 (4)	184	144 (11)	104 [70, 178]	73	0	0	0	–	16	–	–
GammaKnife[a]	10 (1)	30	20 (1)	733 [353, 1107]	50	0	0	0	–	20	–	–
MET total	1,000 (102)	6,840	5,740 (474)	85 [55, 124]	83	97	55	0	–	78	18	582

Records: source-specific patient records with at least one QC-passed scan; parentheses give held-out test counts before task-specific filtering. Visits: kept scans. Pairs: consecutive kept scans of one patient (adjacent intervals). Δt: interval length, median [IQR] over pairs with known dates. History: pairs whose start scan has at least one earlier kept scan of the same patient. 4 seq.: kept scans with T1, T1ce, T2 and FLAIR all present; 4 acq.: all four acquired, not synthesised. GT seg.: kept scans whose 4-class segmentation is the collection's released annotation; the rest are model-predicted. S surgery, R radiotherapy (MET: stereotactic), C chemotherapy or systemic therapy, U treatment of unrecorded type; an interval can contain several classes; – none. U: GBM, SAILOR intervals whose recorded treatment status is 'unknown' (dated); MET, Yale, whose release states that every patient received radiosurgery, whole-brain radiotherapy or resection without type or date, credited to the interval that starts at the baseline scan. [a]GammaKnife days are a course-relative clock (time from a radiosurgery course to its assessment scan), not acquisition dates. [b]IVYGAP2 releases no dates for 80 of its 202 pairs; Δt is over the 122 dated pairs.

Resource size and longitudinal depth. The standardized inventories contain 816 glioma and 1,029 metastasis patient records. After quality control, 800 and 1,000 records retain at least one visit, comprising 2,613 and 6,840 visits, respectively. Of these, 733 and 989 retain at least two visits, and 273 and 775 retain at least three. Two retained visits provide a candidate forecasting pair; three provide a candidate representation-learning triplet or a completed support interval followed by a query. These are depth requirements: the triplet objective also needs reliable order, while dynamics and residual-based adaptation require verified positive intervals (Section 2.1). A record is a source-specific patient or treatment-folder identifier: PROTEAS identifiers are treatment folders. Splits group folders belonging to the same person. Totals count source-specific records.

Cohort coverage and observed change. The resource is released with its full harmonized inventory, and the experiments apply stricter task-specific filters to it. Table S5 reports this released inventory: every scan that passes quality control and every dated adjacent interval (2,613 / 6,840 scans; 1,765 / 5,740 pairs). Table S6 summarizes the pairs with valid segmentations at both endpoints. The analysed forecasting cohorts (Table S7) further require glioma pairs to lie within one treatment window and every visit to pass the imaging checks, which leaves 669 glioma and 986 metastasis records; per-collection counts differ from Table S5 for this reason. Records are source-specific identifiers, grouped by person for the train/test split. Complete imaging denotes T1, T1ce, T2 and FLAIR. Unknown treatment denotes dated SAILOR records with unspecified type, or Yale's cohort-level treatment status assigned to the baseline interval. GammaKnife times measure the interval from a radiosurgery course to its assessment scan.

Patient-cluster bootstrap 95% CIs for growth and regression are 52% [50, 55] and 39% [37, 42] in glioma, and 40% [39, 42] and 40% [38, 41] in metastases. Median absolute lesion-volume changes

Table S6: Tumour change is spatially concentrated and spans growth and regression. Median changed regions occupy 2.3% of the brain in glioma and 0.5% in metastases.

		Direction (%)			Median $\lvert\Delta V\rvert$ (ml)					Change region		$\lvert\Delta V\rvert$ vs Δt	MET lesions (%)	
Collection	Pairs	Grow	Stab.	Shr.	Les.	Oed.	Enh.	Nec.	Cav.	% brain	% $\leq$5 mm	ρ [95% CI]	New	Van.
GBM (stable: $\lvert\Delta V_{\text{lesion}}\rvert \leq 1$ ml)														
UCSF	293	54	19	28	3.1	2.4	0.4	0.0	0.7	0.7	96	$-.41$ [$-.50, -.30$]	–	–
MU	301	56	5	39	21.4	16.1	4.2	0.2	2.3	3.7	76	$-.17$ [$-.29, -.06$]	–	–
LUMIERE	456	54	7	39	14.7	11.9	2.7	1.2	1.7	3.1	77	$-.01$ [$-.11, .09$]	–	–
Burdenko	171	61	2	36	28.4	21.6	6.9	5.2	1.0	5.8	68	$.03$ [$-.12, .18$]	–	–
SAILOR	236	48	9	43	10.1	8.3	2.0	0.6	0.0	1.4	84	$.10$ [$-.10, .30$]	–	–
IVYGAP2	193	44	10	46	9.7	7.2	1.1	0.1	0.0	1.8	93	$-.25$ [$-.45, .02$]	–	–
UCSD	44	66	2	32	27.3	22.2	5.3	0.8	2.7	3.5	74	$-.34$ [$-.57, -.02$]	–	–
RHUH	31	0	0	100	62.0	29.8	18.1	7.5	0.0	5.9	66	$-.09$ [$-.49, .31$]	–	–
All GBM	1,725	52	8	39	13.1	10.8	2.6	0.4	0.4	2.3	82	$-.14$ [$-.20, -.08$]	–	–
MET (stable: $\lvert\Delta V_{\text{lesion}}\rvert \leq 0.15$ ml)														
Yale	5,448	40	20	40	3.5	2.6	0.6	0.0	0.0	0.5	87	$-.14$ [$-.18, -.10$]	42	43
PROTEAS	125	42	16	42	3.9	3.8	0.9	0.0	0.0	0.8	87	$-.16$ [$-.31, .03$]	19	21
All MET	5,573	40	20	40	3.5	2.7	0.6	0.0	0.0	0.5	87	$-.14$ [$-.18, -.10$]	41	43

Adjacent scan pairs of the standard split with valid segmentations at both ends. Direction: % of pairs, lesion (all four classes); 95% patient-cluster bootstrap CI for all GBM / all MET: grow 52 [50, 55] / 40 [39, 42], shrink 39 [37, 42] / 40 [38, 41]. $\lvert\Delta V\rvert$: median absolute volume change per class, ml (lesion, oedema, enhancing, necrosis, cavity). Median $\lvert\Delta V_{\text{lesion}}\rvert$ within change-magnitude quartiles Q1–Q4: GBM 1.5 / 7.7 / 22.3 / 67.5 ml; MET 0.0 / 1.4 / 7.6 / 40.1 ml. Change region: voxels whose 4-class label differs between the two scans; median % of the brain, and median % of changed voxels within 5 mm of the t_0 lesion boundary. ρ: Spearman correlation of $\lvert\Delta V_{\text{lesion}}\rvert$ with Δt, 95% patient-cluster bootstrap CI (pairs with a known interval). New / Van.: % of MET intervals with at least one new or vanished lesion (lesions matched by overlap or centroid distance ≤ 10 mm). RFUDS and GammaKnife are not included in these change summaries.

within quartiles Q1–Q4 are 1.5, 7.7, 22.3 and 67.5 mL for glioma, and 0.0, 1.4, 7.6 and 40.1 mL for metastases.

Harmonization and quality control. Quality control combines automated checks with case-board review; Figure S4 distinguishes unusable acquisitions from correctable processing errors. Versioned visit-level manifests define the observations used for training and evaluation.

Time and treatment information. The record schema separates scan timing, treatment timing, and information availability. Visit indices encode order; verified dates establish elapsed time. Treatment records retain class, documented onset and end, available course details, regimen updates where documented, and explicit missingness. Dated episodes, bracketed events and cohort-level treatment status are stored as distinct record types. Missing interval records in Yale retain an unknown-treatment label. Spatial dose plans retain source provenance; resection maps derived from follow-up images are outcome measurements and are excluded from forecast inputs. Eligibility for conditioning depends on record provenance, including whether a course is supplied retrospectively or documented at the forecast origin (Section 2.1).

Forecasting samples and evaluation populations. A sample consists of a retained current scan, an acquired follow-up, their elapsed interval, supplied treatment information where eligible, and the observed patient prefix. Task manifests distinguish image forecasting, treatment-record conditioning, and history from completed pre-target intervals. Provenance flags distinguish manual from model-generated references and acquired from synthesized sequences, enabling separate evaluation of these groups. A metric case is one modality of an eligible interval. Appendix B defines the metrics, reference predictions, and aggregation protocol.

Reproducibility and access. The data protocol groups each person's visits, lesions, modalities, and history windows before assigning a split. Within the training set, the scheme in Appendix C.11 further separates representation and shared-dynamics learning from fitting the treatment and history edits. The planned release includes preprocessing and quality-control code, source mappings,

derived-data provenance, and task-specific manifests under the terms of each source. Restricted images remain available through their original access procedures.

D.3 Dataset-statistics panels

Figure 4 uses red for glioma and blue for metastases. Each panel selects records according to the measurement it displays; the following denominators therefore differ from the full inventory.

Scan and treatment coverage (a, b). Panel (a) includes scan days and adjacent intervals of at least one day, with densities estimated on $\log_{10}$ time using bandwidth 0.12 and normalized to each row's mode; ticks mark medians. Panel (b)'s imaging-depth display includes 716 glioma and 974 metastasis records with at least two dated scans and one recorded course. Documented resection extent is available for 360 operations. Drug classes can overlap: the glioma drug display covers 494 patients, whereas the metastasis display covers 45 systemic-treatment courses. Dose and fractionation are jointly available for 188 of 677 glioma and 52 of 189 metastasis radiotherapy courses.

Magnitude and tissue composition (c, d). Panel (c)'s volume-ratio scatter uses 1,722 glioma and 4,242 metastasis pairs with positive lesion volume at both endpoints; ratios are clipped to $[1/8, 8]$ for display. Change-category summaries use all 1,725 and 5,573 valid segmented pairs, with absolute stable-change thresholds of 1 mL and 0.15 mL, respectively. Panel (d) uses 1,642 and 4,686 timed pairs with positive baseline volume. Each tissue-volume change is divided by baseline whole-lesion volume. Curves are 5%-trimmed means in interval bins with boundaries 0, 14, 30, 60, 90, 180 and 365 days, retaining bins with at least 30 pairs. Whole-tumour bands are 95% patient-bootstrap intervals from 500 resamples.

Treatment timing and response groups (e, f). Panel (e) includes 2,503 glioma and 6,809 metastasis visits with known relative days, including day zero. Times are measured from the first retained scan; negative course starts precede that scan. Between-class gaps connect the first course to the first later course of the next class, excluding same-day starts; within-class gaps connect consecutive starts. Panel (f) bins relative volume change at -20%, -10%, 10% and 20%, so its central category differs from panel (c)'s absolute-volume definition. Treatment combinations with fewer than 20 pairs are pooled. The sequence display covers 673 glioma and 55 metastasis records with dated courses; same-day events are ordered surgery, radiotherapy, then systemic treatment, and repeated systemic entries are collapsed.

Acquired/synthesized modality and expert/generated annotation flags remain separate through evaluation. Endpoint-derived cavity maps are outcomes; forecast inputs use independently sourced plans (Section C.5).

D.4 Processing and quality control

Figure S3 summarizes the common imaging and record pipeline. Task manifests apply the distinct eligibility rules for representation tuples, timed forecasting pairs and adaptation support (Section 3).

Thumbnails follow one patient and axial slice through steps 1–10; defect examples failed the corresponding check. Dated scan pairs become forecasting candidates subject to positive-interval and provenance checks.

D.5 Image and record harmonization

Scans are resampled to 1 mm isotropic, skull-stripped with SynthStrip (Hoopes et al., 2022) and registered with ANTs (Avants et al., 2011): rigid within visits to T1ce, affine to the patient's baseline, then series realignment to the medoid visit. Brain-mask Dice below 0.92 triggers SyN rescue. A single transform places the series on a brain-centred $192 \times 192 \times 160$ grid. Within-brain intensities map the 1st percentile to 0 and median to 0.5, capped at the 99.5th percentile; visible bias fields receive N4 correction (Tustison et al., 2010).

The BraTS 2025 MissingMRI model synthesizes a missing fourth sequence from the other three, retaining a synthetic flag. Dataset-provided expert masks are retained and harmonized to the study

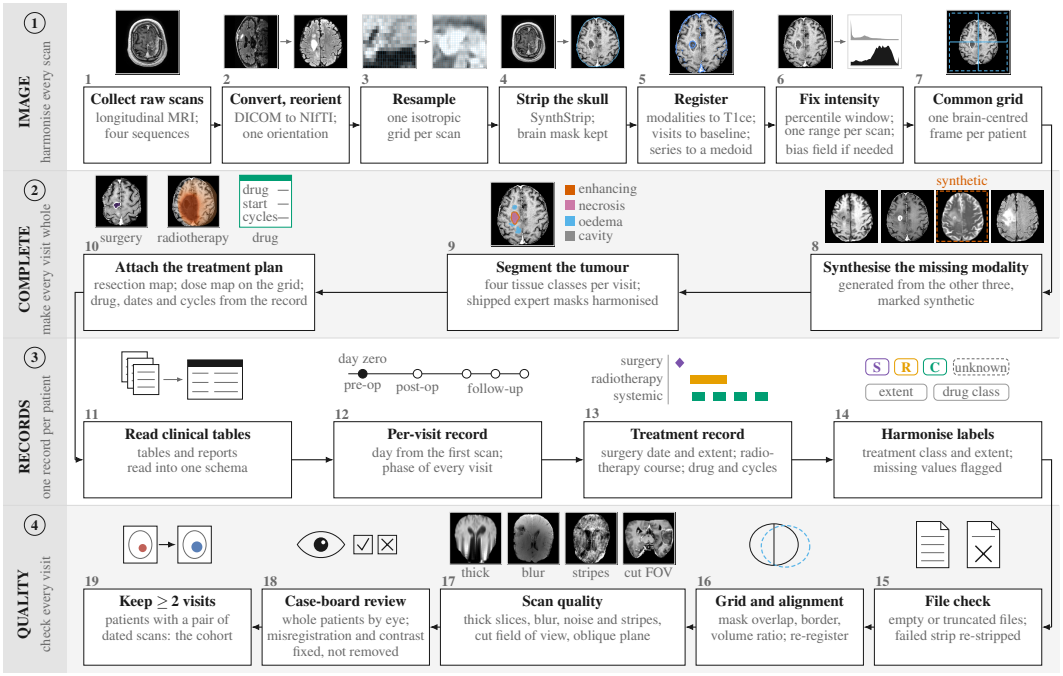

Figure S3: A common imaging grid and linked records turn heterogeneous MRI collections into longitudinal prediction data. Image, Complete, Records and Quality stages produce candidate dated pairs; lanes 2 and 4 run right to left.

label scheme. Where expert annotations are unavailable, the BraTS 2025 post-treatment glioma model (metastasis model for Yale) supplies automated tumour segmentations. Dose maps are resampled to the patient grid, post-operative cavities supply outcome-derived resection maps, and records supply drugs, dates and cycles. Clinical time is anchored to the first scan. Treatment class (surgery, radiotherapy, systemic or unknown), resection extent and drug class retain missingness flags.

Collection-specific settings. Baseline registration is rigid for GK, PROTEAS, IVYGAP2, RFUDS and SAILOR. Yale and PROTEAS arrive skull-stripped; synthesis applies to Burdenko, IVYGAP2, LUMIERE and Yale, and N4 to MU. Unlisted options use library defaults. Forecast-specific grids and input provenance follow Sections C.12 and C.5.

D.6 Quality control and retention

File checks remove visits with brain-mask volume outside 0.80–$1.40\times$ the patient median, Dice to the median visit below 0.78, or saturated fraction above 0.30; failed skull strips are repeated. Grid checks trigger re-registration for Dice below 0.70, border contact above 3%, or baseline volume ratio below 0.65. T1ce quality checks at 2 mm use gradient axis ratio < 0.72 for thick slices (GBM only), sharpness < 0.19 for blur, stripe score $z \geq 1.5$, brain-volume ratio < 0.88 for truncated coverage, and oblique-plane inspection. Case-board review adjudicates borderline visits and whole patients: lesion-clear thick scans can be retained, and correctable alignment or contrast changes are repaired. Cohort pairs require positive dated intervals; GBM pairs also lie within one treatment window. Table S7 gives retention by collection.

Visit accounting. Table S7 counts imaging sessions. Each removed visit is assigned to the first criterion it fails, so each row satisfies raw = removed + retained.

The analysed pair cohorts contain 669 glioma patient records with pairs inside one treatment window and 986 metastasis records with dated pairs. Patient records are source-specific identifiers; PROTEAS counts treatment folders rather than unique people.

Table S7: Analysed forecasting cohorts. The experiments apply stricter filters than the released inventory of Table S5: 65.5% of raw glioma visits and 55.0% of raw metastasis visits enter the analysed pairs. Removals are assigned to the first failed criterion; columns follow Figure S3.

		Scans removed by exclusion criterion											
Collection	Raw	Processing	No mask	No tx. window	Files[†]	Grid[§]	Quality	Review	No dated pair	Retained	Kept (%)	Records	Pairs
Burdenko	645	5	·	7	69	101	101	3	80	279	43.3	96	175
LUMIERE	638	·	·	·	1	58	13	8	5	553	86.7	84	466
MU	596	·	·	104	10	·	65	·	82	335	56.2	113	210
UCSF	596	·	·	18	·	4	2	·	4	568	95.3	284	284
IVYGAP2	390	3	·	91	1	98	20	·	9	168	43.1	21	139
SAILOR	270	·	·	21	·	·	·	·	·	249	92.2	27	174
UCSD	184	·	·	14	·	·	·	·	107	63	34.2	26	37
RHUH	120	·	·	26	40	·	9	·	9	36	30.0	18	18
Glioma total	**3,439**	**8**	·	**281**	**121**	**261**	**210**	**11**	**296**	**2,251**	**65.5**	**669**	**1,503**
Yale	11,877	475	2,591	–	907	1,011	210	106	142	6,435	54.2	894	5,451
RFUDS	257	1	·	–	55[†]	·	·	2	15	184	71.6	39	144
PROTEAS	186	·	·	–	1	·	3	4	4	174	93.5	43	125
GK	76	·	·	–	·	·	1	1	44	30	39.5	10	20
MET total	**12,396**	**476**	**2,591**	**–**	**963**	**1,011**	**214**	**113**	**205**	**6,823**	**55.0**	**986**	**5,740**

Raw and retained counts are scans; records are source-specific patient records; pairs are adjacent visits. Tx.: treatment. [†] RFUDS: 3 file-check failures and 52 visits absent from the QC manifest. [§] Grid checks count unique visits. – Metastasis pairs are not restricted to treatment windows. · = 0.

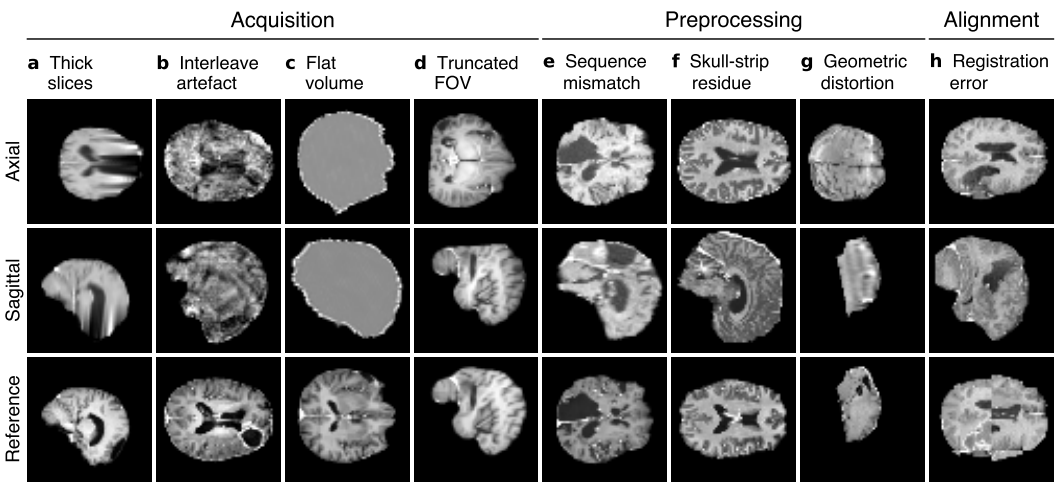

Figure S4: Serial MRI requires quality control within and between visits. **(a–d)** Thick slices, stripes, absent anatomy and truncated coverage. **(e–g)** Sequence mismatch, residual skull/dura and resampling distortion. **(h)** A 12 mm inter-visit offset. Top rows: axial and sagittal views; bottom: a comparison visit or checkerboard. The illustrated cases a–d and g are excluded; e, f and h are corrected.

The released inventory retains 2,613 / 6,840 scans of 800 / 1,000 patient records (Table S5); the analysed forecasting cohorts below apply the stricter task filters of Appendix D.2.

Retention reflects longitudinal eligibility. Missing segmentations are Yale's largest listed exclusion (2,591 visits); lack of a pair excludes 107 UCSD visits. In glioma, pair and treatment-window requirements exclude 296 and 281 visits, respectively, each exceeding the 210 quality exclusions. Retention thus reflects longitudinal eligibility as well as image quality.

D.7 Observed error modes

Figure S4 shows representative cases from visual review. Thick-slice review removes 177 visits and retains 87 whose lesions are delineable in all three views; Burdenko, MU and IVYGAP2 contribute

90, 59 and 14 removals, and other collections 14. Visual review determines retention after the axis-ratio flag. Thin-slice metastasis collections have no thick-slice rule. Stripe artefacts remove 15 glioma and 112 metastasis visits; truncated fields remove 17 and 81, structureless volumes 1 and 21, and whole-series exclusions for truncation or distortion total 11 and 74.

Corrections preserve usable trajectories: 7 glioma visits receive intensity correction for sequence changes; 5 lose a pre-contrast file duplicating the post-contrast image. Of 44 alignment flags, 27 are confirmed and rigidly re-registered (brain-mask Dice ≥ 0.93 afterwards), while 17 reflect mask differences from truncated coverage. Eleven brain masks are tightened: 7 mild rims with 11–15% excess volume and 4 gross failures. Series-level review distinguishes correctable processing errors from unusable acquisitions.

E LATENT-SPACE ANALYSIS DETAILS

Figure 2 analyzes GBM observations encoded by the final frozen change-sensitive autoencoder used by LMWM, trained with the reconstruction and change-direction objectives of Section 2.2. The baseline-to-follow-up difference is $\Delta z = z(t_k) - z(t_b) \in \mathbb{R}^{4\times32\times32\times32}$. The image-space comparator ΔI is the corresponding image difference averaged over 4^3 blocks onto the same spatial grid. Square brackets in the text denote 95% patient-bootstrap confidence intervals.

Cross-modal agreement and acquisition perturbations. Panel (a) measures Spearman correlation of per-cell change energy between T1 and contrast-enhanced T1 inside the tumour. The 700 points are follow-ups; colour marks the representation with higher agreement. The latent-minus-image difference is $\Delta r = +0.056$ [$+0.042$, $+0.069$], and 70% of points lie above the diagonal. The $\Delta z/\Delta I$ ratios of R^2 are $1.10\times$ for Dice, corresponding to an absolute gain of $+0.056$ [$+0.031$, $+0.082$], and $1.09\times$ for $\Delta \log V$, corresponding to $+0.033$ [-0.010, $+0.077$]. The response to a $\pm5\%$ bias field is smaller by a factor of 4.8, and far-field drift by factors of 2–4.9.

Disease state and preceding change. Panel (b) relates log latent-change energy within the tumour to baseline tumour size over 1,742 follow-ups. Colour encodes the preceding interval's latent-change magnitude; grey marks observations without a preceding interval. Partial Spearman correlations after removing cohort and $\log \Delta t$ are 0.69 for tumour size and 0.58 for preceding latent change. Adding baseline tumour size, shape, and location to a model containing cohort, elapsed time, and modality increases out-of-fold R^2 from 0.41 to 0.70, a gain of $+0.29$ [$+0.23$, $+0.37$].

Temporal scaling and shared history. Panel (c) uses 244 segments from 155 patients and normalizes the plotted magnitude to the ≤ 30-day bin. The dashed random-walk reference follows the plotted linear-in-time relation; the observed log–log slope is $\beta = 0.10$. Annotations give normalized off-diagonal Gram entries (%) between within-trajectory changes, grouped by shared time. Overlap increases with shared time ($\rho = 0.52$). The step-plus-random-walk model gives the highest held-out Gram-matrix R^2 (0.38; inset).

Treatment-aligned volume and latent changes. Panel (d, top) plots enhancing-tumour (ET) volume change relative to the last pre-treatment scan against days since treatment start. Chemotherapy contributes 1,940 visits from 95 patients, radiotherapy 1,112 visits from 115 patients, and surgery 703 visits from 95 patients; 94% of the surgery data come from LUMIERE. The chemotherapy curve is a smoothed local median, implemented as the weighted mean of the 40th–60th percentile slice with span 0.3. Radiotherapy and surgery use tricube LOESS means after winsorizing values at the 5th–95th percentiles. Bands are 68% patient-bootstrap intervals, purple marks overlap between the chemotherapy and radiotherapy bands, and rings mark curve nadirs.

Panel (d, bottom) identifies windows outside a record-free null band after residualizing on visits without a recorded course from the cohorts in which each treatment is identifiable. Chemotherapy is associated with a lower share of latent-change energy in ET (mean -0.39 SD during days 14–150), whereas radiotherapy is associated with greater latent-change magnitude (peak $+0.50$ SD during days 30–90). No corresponding surgery window is detected. SD denotes the standard deviation over record-free visits. Temporal windows are not corrected for multiple comparisons.

E.1 Consistent maps across visits

Figure S5 shows the consistent map H_i (Section 2.5) beside signed scan-to-scan changes in six longitudinal series.

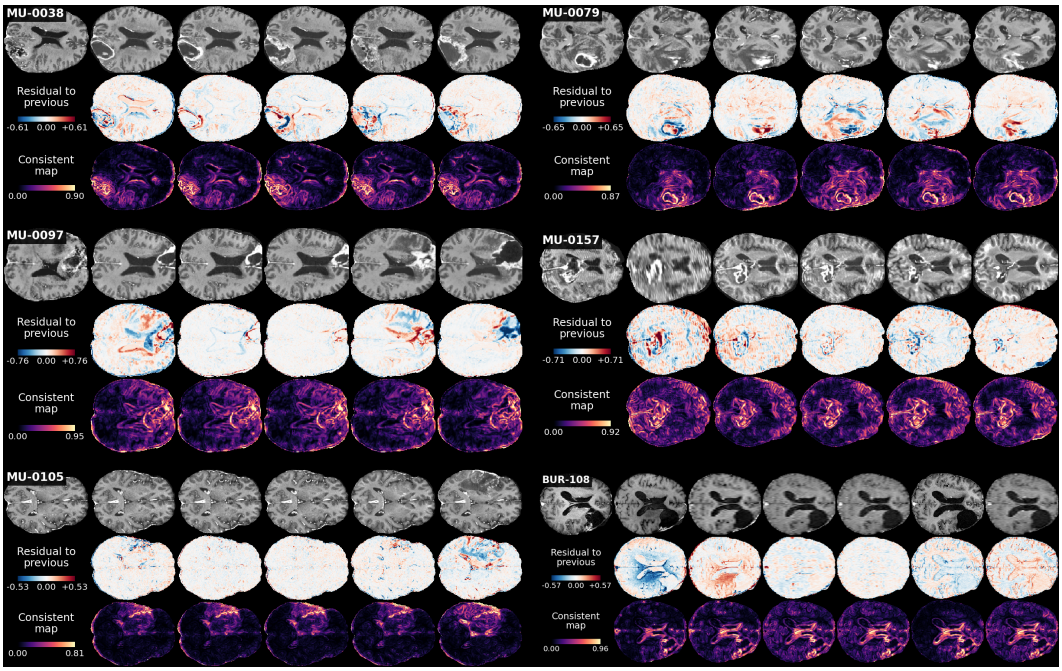

Figure S5: **Consistent maps and signed image changes across visits.** Six glioma series from MU-Glioma-Post and Burdenko. Each case shows registered contrast-enhanced T1 scans (top), signed differences from the preceding visit (middle; red: brighter, blue: darker) and the magnitude of the consistent map updated at that visit (bottom). Visits run from left to right on the same axial slice. Differences and maps each use one colour scale per patient; neither is shown at baseline.

F Change direction versus state direction

The representation objective of Section 2.2 supervises the direction of latent change rather than the next latent state. Write the auxiliary affine map as $g(z_i, d_i^{\text{prev}}) = Az_i + Bd_i^{\text{prev}} + b$. Predicting the next state by adding the current state gives $(A + I)z_i + Bd_i^{\text{prev}} + b$, which is another affine map with exactly the same parameter count. Moreover,

$$\|g - (z_{i+1} - z_i)\|_2^2 = \|z_i + g - z_{i+1}\|_2^2. \tag{36}$$

The squared-error identity holds for every encoder; cosine supervision differs because adding z_i changes directions. For example, in two spatial coordinates of one latent channel, let $z_i = (100, 1)$, $z_{i+1} = (100, 2)$ and the predicted state be $(100, 0)$. State cosine similarity is approximately 0.9998, while observed and predicted changes have cosine -1. Decoded magnitude is evaluated separately because cosine is scale-invariant away from its numerical floors.

G Extended related work

World models pair an observation representation with a learned state transition and use it to predict the consequences of actions (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025).

Prediction targets and available information. Medical image generators differ in the target they produce and the information available at inference. CardiacAging learns age-conditioned cardiac MRI synthesis from cross-sectional data using a conditional GAN (Campello et al., 2022), while

CounterSynth generates label-conditioned, diffeomorphic brain-image transformations for counterfactual data augmentation (Pombo et al., 2023). DiffuseMorph learns deformable registration from a supplied moving–fixed image pair; its target image is an inference input (Kim et al., 2022). Longitudinal forecasting instead predicts an unobserved follow-up from earlier observations. SADM conditions diffusion on prior images through spatial and temporal attention (Yoon et al., 2023), and TADM generates follow-up image differences conditioned on baseline MRI and elapsed time, with an auxiliary brain-age objective (Litrico et al., 2024). MambaControl combines Mamba-based diffusion and control networks with Fourier-enhanced anatomical graph features for follow-up MRI prediction (Yang et al., 2025a). Section 4 compares the forecasting methods on the same target and test cases with matched time and treatment encodings.

What should a longitudinal representation learn? Δ-LFM trains longitudinal representations with ArcRank, which aligns within-patient latent orientations and ranks latent magnitudes by chronology, before learning temporal latent flow (Chen et al., 2026). ArcRank organizes observed trajectories. The objective of Section 2.2 instead asks the encoder to make the next change predictable.

How should patient history change a forecast? BrLP conditions latent diffusion on regional brain volumes predicted by a disease-course model, with a longitudinal variant that fits this model to the patient's volume history (Puglisi et al., 2024). ImageFlowNet evolves multiscale image representations through continuous-time flow fields (Liu et al., 2025a), and its extended version also fine-tunes these flow fields using patient history. Our design adapts the shared transition without per-patient fitting: completed forecast errors are summarized into a coherence-weighted map and read through a gated attention branch. LEADS decomposes dynamics into shared and environment-specific terms (Yin et al., 2021); ALPaCA performs Bayesian linear regression on learned features (Harrison et al., 2018). We keep the shared-plus-correction structure but replace the coefficient solve with a fixed summary of the errors. Additive adaptation of a frozen backbone follows adaLN-Zero, ControlNet and gated cross-attention (Peebles & Xie, 2023; Zhang et al., 2023; Alayrac et al., 2022). Here it makes the treatment-agnostic transition an exact special case of the adapted model, which turns it into a built-in reference.

Treatment-conditioned disease dynamics. TaDiff jointly generates future MRI and tumour segmentations using prior MRI, treatment type, and treatment-day information (Liu et al., 2025b). GlioPREDICT (3D-GlioPREDICT) uses a pretrained MAISI autoencoder and a ControlNet on a frozen latent rectified-flow backbone, conditioned on baseline MRI, spatial radiotherapy dose, and follow-up time (Belkacemi et al., 2026). MeWM combines action-conditioned 3D liver CT generation with treatment planning (Yang et al., 2025b). In our design, the treatment record enters as a spatiotemporal response field that modulates every block. A gated history branch processes the treatment text together with a map of errors from the patient's completed forecasts. Both components modulate the same frozen transition. Section 4 compares these methods under identical time and treatment encodings.