OncoFlow: Context-Conditioned Flow Matching for Brain MRI Forecasting

Anonymous authors
Paper under double-blind review

Abstract

Forecasting how brain tumors will evolve on MRI could enable earlier assessment of disease progression and treatment response. This remains challenging because tumor evolution is rapid and heterogeneous, shaped not only by prior imaging but also by treatment interventions and underlying tumor biology. Existing methods often fail to capture these clinically meaningful changes, while conventional evaluation metrics are dominated by stable anatomy and can therefore reward near-identity predictions. We introduce OncoFlow, a clinically conditioned 3D flow-matching framework for patient-specific MRI forecasting. Its source distribution spatially interpolates between the latest MRI and Gaussian noise, progressively reducing dependence on the observed anatomy as the forecast horizon increases. A multipath velocity network integrates visit-aligned longitudinal imaging memory with global conditioning on treatment, genomics, forecast horizon, and generative time. To evaluate whether predicted scans capture disease evolution rather than merely reconstruct static anatomy, we introduce $R_{pattern}$, a Residual Pattern Alignment metric that quantifies the spatial fidelity of predicted changes. Across three longitudinal brain-tumor datasets, OncoFlow achieves state-of-the-art overall performance, consistently improving change fidelity while remaining competitive in whole-image reconstruction. Qualitative and ablation analyses show that soft source conditioning and horizon-dependent decay suppress spurious changes in healthy tissue without collapsing to near-identity forecasts. Crucially, conditioning on treatment and genomic information further improves forecasting performance over imaging history alone, demonstrating that non-imaging clinical variables provide complementary signals about future tumor evolution. Together, these findings establish clinically conditioned generative modeling as a promising route towards forecasts that preserve patient anatomy while remaining sensitive to meaningful, treatment- and biology-driven change.

1 Introduction

Brain tumors exhibit highly heterogeneous progression and treatment responses, making it difficult to determine whether a therapy is working for an individual patient. The answer often cannot be determined until weeks or months later, when the next MRI scan is acquired (Wen et al., 2017). During this waiting period, tumors can evolve unpredictably, sometimes progressing aggressively and narrowing the patient's window for effective intervention, with consequences for both outcomes and healthcare costs (National Cancer Institute, 2024). A model that predicts a patient's future MRI from their existing data could provide an earlier estimate of disease progression, shifting clinical decision-making from reactive monitoring to proactive and personalized intervention.

Longitudinal generative models have incorporated imaging sequences (Yoon et al., 2023) and temporal intervals (Litrico et al., 2024) to predict subsequent examinations. However, imaging history alone does not explicitly identify the interventions associated with an observed trajectory or planned before the next scan; which is a consequential omission in oncology, where progression reflects both disease biology and treatment response. TaDiff (Liu et al., 2025b) advances this direction through treatment-conditioned diffusion. However, it does not associate each image with the treatment that preceded it, and it does not model tumor biology

A complementary issue concerns the starting point and structural guidance of generation. Noise-initialized diffusion forecasts must reconstruct patient anatomy solely through conditioning, whereas methods like SDEdit (Meng et al., 2022), RePaint (Lugmayr et al., 2022), and ControlNet (Zhang et al., 2023) mix noise with reference images or structural priors to enforce spatial consistency. More recently, CRONOS (Disch et al., 2026) introduces flow matching from observed sequences to provide a direct anatomical anchor. However,

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.

these formulations rely on a static prior or fixed conditioning strength that does not dynamically adapt across forecast horizons. Neither pure noise reconstruction, static reference mixing, nor transport from observations explicitly determines how spatial anchoring should decay or adjust as the forecast extends further into the future.

These gaps motivate two questions: (1) *How should a generative source balance spatial preservation and stochastic freedom across prediction horizons?* and (2) *how should a generator combine visit-specific evidence with the biological and treatment context of a future forecast?*

We address the first question by treating the generative source as a forecast-dependent modeling choice. Noise-based generation and transport from an observed scan (Disch et al., 2026) offer different starting assumptions about how strongly the future should inherit the present. Yet the relevance of an observation changes with the interval being forecast. We therefore reformulate conditional flow matching (Lipman et al., 2023) around a spatially varying mixture of the latest scan and Gaussian noise, whose reliance on the observation decays with the forecast horizon. The horizon thus determines not only what the velocity network predicts, but also the distribution from which its trajectory begins. This creates a continuous family of forecasting problems between observation-anchored transport and stochastic generation, with the balance adapting across space and clinical time.

We address the second question through complementary conditioning pathways. A longitudinal memory bank retains prior visits and associates each acquisition with its acquired clinical information. A separate global pathway supplies genomic features, the specified future treatment, the forecast horizon, and generative time. This separation distinguishes evidence about the patient's observed trajectory from the context under which a future image is generated. Together, the pathways support patient-specific forecasting from imaging history and molecular and treatment information, a combination that remains underexplored in longitudinal MRI generation. Treatment is used as observational conditioning, without interpreting alternative treatment settings as causal effects.

Our contributions are:

- **Spatial and horizon-dependent source formulation.** We introduce a spatially varying, horizon-decaying source for flow matching that interpolates between the latest observation and Gaussian noise, making the forecast interval part of the generative starting distribution.
- **Longitudinal and global clinical conditioning.** We develop a multipath velocity network that combines visit-aligned longitudinal memory with genomic, treatment, and temporal conditioning.
- **Change-aware evaluation.** We show that image-similarity metrics can favour near-identity forecasts, and introduce R_{Pattern} with a change-magnitude ratio to evaluate predicted change. Under this evaluation, OncoFlow gives the largest change-fidelity gains on cohorts with pronounced progression.

2 Related Work

Longitudinal image generation. Predicting future individual anatomy relies on diverse inputs and generative architectures. Multi-acquisition sequence models use sequential temporal contexts via transformers (SADM (Yoon et al., 2023)) or temporal recurrent GANs (TR-GAN (Fan et al., 2022)) to project multi-session dynamics. Conversely, single-scan baselines either model intensity residual changes directly (TADM (Litrico et al., 2024)) or integrate external spatial and clinical priors into latent diffusion frameworks (BrLP (Puglisi et al., 2024)). Despite leveraging spatio-temporal dynamics, early sequence- and difference-based models often omit explicit molecular covariates and clinical factors, limiting their predictive fidelity in complex settings such as oncology.

Treatment-aware tumor forecasting. TaDiff (Liu et al., 2025b) advances clinically informed MRI forecasting by conditioning diffusion generation on longitudinal imaging and treatment information. Nevertheless, it does not incorporate measured genomic features, leaving this source of between-patient variation without explicit representation. Furthermore, when observing the patient's history, the model is not explicitly informed which treatment preceded each MRI, leaving the association between therapeutic interventions and subsequent imaging findings implicit. Training also required approximately 350-GPU hours on a single NVIDIA V100 (Liu et al., 2025b), incurring a substantial computational cost that limits its retraining and inclusion in our benchmark comparison. These limitations motivate efficient forecasting models that combine flexible longitudinal memory with explicit visit-level imaging–treatment associations and global genomic and future-treatment conditioning.

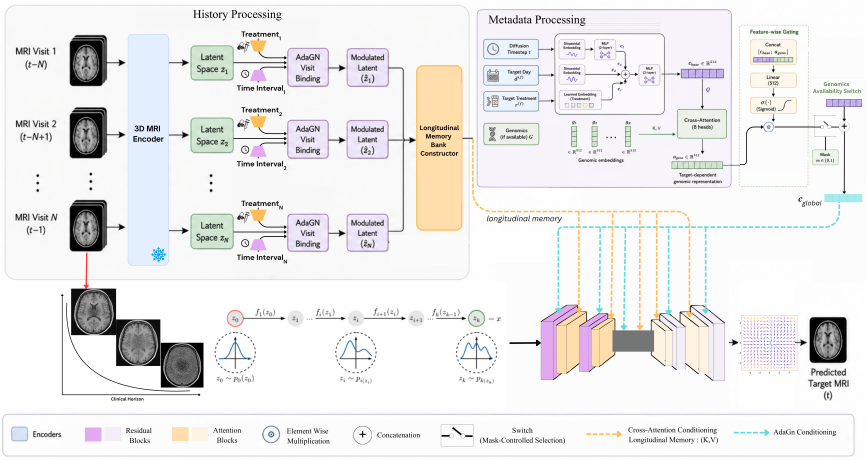

Figure 1: **OncoFlow architecture.** Visit-conditioned MRI histories form a longitudinal memory, while target metadata and optional genomics condition a 3D flow-matching backbone. Spatial conditioning and horizon decay preserve baseline anatomy while forecasting tumor evolution at the target clinical horizon.

Continuous-time and flow-based image forecasting. Continuous-time frameworks model longitudinal anatomical evolution without rigid interval constraints. ImageFlowNet (Liu et al., 2025a) evolves patient trajectories via neural ODEs/SDEs in latent space, but relies on abstract dynamics that could fail to preserve precise local spatial anchors and it lacks explicit clinical or genomic conditioning. TumorFlow (Biller et al., 2026) integrates biophysical reaction-diffusion PDEs to constrain continuous growth, yet it depends on hand-crafted mechanical assumptions that generalize poorly to complex treatment responses or global temporal dynamics. In contrast, conditional flow matching (Lipman et al., 2023) regresses velocity fields between source and target distributions. CRONOS (Disch et al., 2026) adapts this for sequence-to-image transport by broadcasting context volumes uniformly across the field of view; however, this spatially static, horizon-agnostic source fails to decay as the forecast window widens and ignores localized tumor occupancy, often over-smoothing localized anatomical changes. We resolve these limitations by formulating a spatially and temporally decaying source derived from tumor occupancy, supported by a multipath velocity network that integrates longitudinal memory alongside visit-aligned genomic, treatment, and temporal covariates.

3 Methodology

We consider forecasting a future MRI from a patient's imaging history and clinical context. Let $\mathcal{H} = \{(x_n, \tau_n, a_n)\}_{n=1}^{N}$ denote the observed visits, where x_n is a scan, τ_n its acquisition day relative to diagnosis, and a_n the treatment associated with it. Given the target acquisition day τ_{fut} and a specified target treatment a_{fut}, the model generates x_{fut} such that the prediction horizon is $\Delta = \tau_{\text{fut}} - \tau_N \geq 0$. Patient-level genomic measurements g provide additional conditioning when available.

Our formulation connects three decisions. A spatial and temporal preservation map determines how much trust should be preserved to the previous scan. A clinically conditioned U-Net uses the longitudinal history to predict the transformation from that source to the future scan. Finally, a curriculum gradually increases the contribution of tumor-associated reconstruction errors. We first express these components through a common velocity-regression objective.

3.1 Preliminaries: conditional flow matching

Conditional flow matching (Lipman et al., 2023) learns a time-dependent velocity field that transports a source distribution to a target distribution. Given a source sample x_0, a target sample x_1, and an independently sampled time $t \sim \mathcal{U}[0,1]$, we define the linear path and its velocity as

$$x_t = (1-t)x_0 + t\,x_1, \qquad u = \frac{\mathrm{d}x_t}{\mathrm{d}t} = x_1 - x_0. \tag{1}$$

A neural network $v_\theta(x_t, t, \mathcal{C})$, conditioned on inputs $\mathcal{C}$, learns to predict this velocity by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{x_0, x_1, \mathcal{C}, t}\left[\|v_\theta(x_t, t, \mathcal{C}) - (x_1 - x_0)\|_2^2\right]. \tag{2}$$

At inference, samples are generated by integrating the learned velocity field from $t = 0$ to $t = 1$ with an ODE solver, starting from a fresh source sample.

Two components of this formulation are left open by the framework and are specified by the method: the source distribution from which x_0 is drawn, and the construction of the conditioning $\mathcal{C}$. The source is commonly taken to be an isotropic Gaussian, but the derivation only requires that it can be sampled at inference time; the marginal consistency of the learned field holds for any such choice, including couplings in which the source depends on observed data. We exploit this freedom in Section 3.2 and describe the conditioning network in Section 3.3.

3.2 Spatial and temporal source construction

Two extreme sources and their failure modes. Standard flow matching transports samples from Gaussian noise, $x_0 \sim \mathcal{N}(0, I)$, to the data distribution. In longitudinal prediction, this source gives the model complete freedom: no part of the output is anchored, so lesion appearance, growth, and regression can all be synthesized without constraint. The cost is that the entire image must be produced from noise, including the large majority of the volume that does not change between visits.

The opposite extreme initializes the flow from the most recent scan, $x_0 = x_{\mathrm{prev}}$, providing a direct anchor for observed anatomy. When combined with optimal-transport couplings, this source favours short paths further motivating conservative, small-change predictions. Such predictions can achieve high image-similarity scores while largely reproducing the previous lesion. This emphasis on preservation is poorly suited to glioblastoma progression, where clinically meaningful changes can be large.

Neither extreme matches the structure of the task, predictability varies across the image, and it varies with how much time has elapsed. We therefore construct a source that interpolates between the two extremes, through a preservation map $P(r) \in [0, 1]$ that specifies how much of the observed anatomy each location inherits and how much stochastic freedom it receives.

Soft-dilation spatial preservation. Lesion boundaries are not sharp divisions between predictable and unpredictable tissue; infiltration, edema, and partial-volume effects make the surrounding region progressively less reliable as an anchor. We therefore convolve the lesion mask of the most recent, *already acquired* scan with a normalized box kernel $K_{\mathbf{k}}$ to obtain a soft lesion-influence map:

$$R(r) = (K_{\mathbf{k}} * M_{\mathrm{prev}})(r), \qquad R(r) \in [0, 1], \tag{3}$$

which softly extends lesion influence into surrounding tissue. Spatial preservation is then

$$P_{\mathrm{sp}}(r) = 1 - R(r)^\gamma, \qquad \gamma > 0, \tag{4}$$

where γ shapes the response to intermediate occupancy: values below one broaden the influence of sparse lesion presence, while values above one concentrate reduced preservation in dense neighborhoods. Equation 4 determines *where* the observed anatomy anchors the source, but it is silent on *for how long* that anchoring should hold, and the two questions are not separable.

Horizon-dependent preservation decay. At long horizons, even tissue far from the lesion becomes a less reliable guide as distant recurrence, mass effect, ventricular change, and treatment-related white-matter changes can accumulate over months to years in regions beyond the lesion mask, so a source that anchors them fully would constrain the model against changes that genuinely occur.

We therefore attenuate spatial preservation by a factor that decays with the clinical interval Δ,

$$P_\Delta(r) = \left(1 - R(r)^\gamma\right) \exp(-\lambda_{\mathrm{bg}}\Delta), \qquad \lambda_{\mathrm{bg}} \geq 0. \tag{5}$$

Preservation now accounts for elapsed time: shorter intervals favor stronger reliance on the observation, while longer intervals gradually relax this anchor to accommodate greater uncertainty.

A reference horizon is specified $h_{\mathrm{ref}} > 0$ and a retained fraction $\rho \in (0, 1]$. Requiring $f(h_{\mathrm{ref}}) = \rho$ gives

$$\lambda_{\mathrm{bg}} = -\frac{\log \rho}{h_{\mathrm{ref}}}, \qquad e^{-\lambda_{\mathrm{bg}}\Delta} = \rho^{\Delta/h_{\mathrm{ref}}}. \tag{6}$$

Thus, each reference interval retains the same fraction ρ of the remaining preservation weight. We fix these parameters based on established clinical statistics (Milano et al., 2010; Jiang et al., 2020).

Source construction. The preservation map combines the last observation with independent noise as $\epsilon \sim \mathcal{N}(0, I)$,

$$x_0(r) = P_\Delta(r)\, x_{\text{prev}}(r) + \sqrt{1 - P_\Delta(r)^2}\, \epsilon(r). \tag{7}$$

The two extremes of Section 3.2 are recovered exactly: $P_\Delta = 1$ gives the observation, and $P_\Delta = 0$ gives the standard Gaussian source. The square-root weighting makes the squared coefficients sum to one, preserving variance for independent, unit-variance components and avoiding the variance reduction of a convex blend. The map is computed once per sample and held fixed throughout the ODE solve. While this initialization uses only the most recent scan, the velocity network (Section 3.3) is additionally conditioned on earlier observations and clinical context to guide the subsequent evolution.

3.3 Clinically Conditioned Longitudinal U-Net

We parameterize the velocity field with a U-Net whose residual and attention blocks receive complementary conditioning signals. Historical scans first acquire their visit-specific clinical context, after which a spatio-temporal module organizes them into a memory bank. The network then combines access to this memory with a direct recent-scan reference and a global condition describing the requested examination.

Longitudinal Memory OncoFlow constructs a longitudinal memory bank by conditioning each historical scan on its clinical context and integrating the resulting representations across visits. For visit n, a sinusoidal day embedding E_{day} and a learned treatment embedding E_{trt} form the visit condition

$$c_n = \text{MLP}_{\text{visit}}\left([E_{\text{day}}(\tau_n); E_{\text{trt}}(a_n)]\right) \in \mathbb{R}^{d_c}, \tag{8}$$

where $[;]$ denotes concatenation and d_c is the condition embedding dimension. This condition modulates the historical image representation F_n through adaptive group normalization:

$$\widetilde{F}_n = \gamma_{\text{visit}}(c_n) \odot \frac{F_n - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}} + \beta_{\text{visit}}(c_n). \tag{9}$$

Here, μ_n and σ_n^2 denote the mean and variance within each channel group, ϵ ensures numerical stability, and $\odot$ denotes elementwise multiplication. The condition-dependent scale $\gamma_{\text{visit}}(c_n)$ and shift $\beta_{\text{visit}}(c_n)$ are predicted by learned mappings, associating each scan with its acquisition time and treatment before temporal modeling.

Writing $S_h = |\Omega_h|$ and suppressing the batch dimension, the conditioned features are arranged into a tensor in $\mathbb{R}^{N \times S_h \times d_h}$ and processed by 4 layers of spatio-temporal transformer blocks. Within each block, spatial self-attention is applied independently to each visit to model intra-scan spatial dependencies. To reduce computational cost, attention is computed on adaptively downsampled feature maps before being restored to their original resolution. Temporal self-attention then relates visits at corresponding patches, followed by a position-wise feed-forward module with a residual update. These operations follow the attention framework of Yoon et al. (2023), modeling both spatial structure and longitudinal relationships.

Variable-length histories are padded within each batch, and a temporal attention mask excludes padded visits. The refined features retain both visit and spatial axes, which are flattened to form the longitudinal memory:

$$\mathcal{M}_{\text{long}} \in \mathbb{R}^{(NS_h) \times d_h}. \tag{10}$$

This memory retains distinct historical features together with their clinical context, allowing the U-Net to retrieve relevant evidence through the resolution-dependent conditioning.

Multi-Path Conditioning The velocity network uses the historical evidence and target clinical context through two complementary paths. At each U-Net attention block, current features query $\mathcal{M}_{\text{long}}$ through cross-attention. Because queries depend on the current features, different regions and resolutions can retrieve different parts of the patient's history.

In parallel, a global path specifies the requested clinical context and the position along the flow. Sinusoidal embeddings of t and Δ, together with the target-treatment embedding, define

$$c_{\text{base}} = \text{MLP}_{\text{global}}\left([E_{\text{flow}}(t); E_{\text{target}}(\Delta); E_{\text{trt}}^{\text{target}}(a_{\text{fut}})]\right) \in \mathbb{R}^{d_g}. \tag{11}$$

This condition describes both the position of the current state along the flow and the clinical examination being requested.

When genomic measurements are available, each genomic variable is projected into a token of dimension d_g. The base condition queries these tokens by cross-attention to obtain a_{geno}. A feature-wise gate determines the genomic update:

$$q_g = \sigma\left(W_g[c_{\text{base}}; a_{\text{geno}}] + b_g\right),$$
$$c_{\text{global}} = \begin{cases} c_{\text{base}} + q_g \odot a_{\text{geno}}, & \text{if genomic measurements are available,} \\ c_{\text{base}}, & \text{otherwise.} \end{cases} \tag{12}$$

For patients without genomic measurements, the genomic update is bypassed such that $c_{\text{global}} = c_{\text{base}}$. The resulting global condition is then injected into every residual block through AdaGN.

Together, memory cross-attention retrieves spatially relevant history, and residual modulation communicates global clinical context throughout the network. These signals specify how $\mathcal{C}$ enters $v_\theta(x_t, t, \mathcal{C}, \Delta)$; their effect is not restricted to a single global embedding. The remaining design choice is how to train this conditioned field to balance image-wide fidelity with tumor detail.

3.4 Curriculum Optimization and Inference

To mitigate severe spatial class imbalance between small tumor regions and extensive healthy brain tissue, we introduce a subject-normalized loss weighting scheme guided by a curriculum schedule. Target follow-up segmentations L_{fut} are strictly leveraged to compute spatial loss weights during training and are not required at inference.

We first construct a continuous, smoothed spatial priority map $B_{\text{fut}}(r)$ over voxel locations $r \in \Omega$:

$$A_{\text{fut}}(r) = \sum_j L_{\text{fut},j}(r), \qquad B_{\text{fut}}(r) = \left[K_{\text{loss}} * \left(A_{\text{fut}} e^{-A_{\text{fut}}}\right)\right](r), \tag{13}$$

where K_{loss} represents a box kernel with constant entries $1/d_K$ ($d_K > 0$). The bounded mapping $x \mapsto xe^{-x}$ saturates multi-label overlaps, preventing arbitrary weight growth. Spatial convolution with K_{loss} extends emphasis into adjacent peritumoral tissue, constructing a soft boundary margin to effectively supervise localized deformations.

To prevent optimization from prematurely collapsing into local minima, an epoch-dependent parameter $\alpha(e) \in [0, 1]$ governs a curriculum transition from global alignment ($\alpha(e) \to 0$) to boundary refinement ($\alpha(e) \to 1$). The raw, mean, and normalized spatial weights are defined as:

$$W_e(r) = 1 + \alpha(e) B_{\text{fut}}(r), \qquad \overline{W}_e = \frac{1}{V} \sum_{q \in \Omega} W_e(q), \qquad \widetilde{W}_e(r) = \frac{W_e(r)}{\overline{W}_e}. \tag{14}$$

The unit base offset guarantees non-zero supervision across all anatomical regions. Subject-wise normalization by the spatial mean $\overline{W}_e$ maintains a mean weight of 1.0 across all subjects, preserving intra-subject spatial prioritization while decoupling tumor volume from gradient magnitudes to stabilize batch optimization.

The overall epoch-dependent training objective scales the point-wise reconstruction loss $\ell_\theta(r)$:

$$\mathcal{L}_e(\theta) = \mathbb{E}\left[\frac{1}{V} \sum_{r \in \Omega} \widetilde{W}_e(r) \ell_\theta(r)\right]. \tag{15}$$

Curriculum schedule. We train with standard MSE before introducing additional tumor emphasis. For start epoch e_s and ramp duration $E_r > 0$, define

$$s(e) = \text{clip}\left(\frac{e - e_s}{E_r}, 0, 1\right), \qquad \alpha(e) = \frac{1 - \cos(\pi s(e))}{2}. \tag{16}$$

At $\alpha = 0$, all weights equal one and Equation 15 reduces to the standard flow-matching objective. The additional tumor contribution then increases smoothly, reaching the full normalized weighting at $\alpha = 1$. An empty target tumor mask gives standard MSE throughout.

3.5 Data and Experimental Design

Datasets. We evaluate our framework across three distinct longitudinal glioma cohorts offering complementary clinical and temporal characteristics. Across all cohorts, datasets were filtered to include only patients with at

least two longitudinal scans. We exclusively use the FLAIR modality for all experiments. All MRI volumes were coregistered, and intensity values were min-max scaled to $[0, 1]$ independently per visit to prevent leakage. Models were trained on a single 40GB GPU for 24 hours on the high resolution images. Further architectural details and training hyperparameters are provided in the Supplementary Material.

MU-Glioma (Mahmoud et al., 2025) features extensive longitudinal coverage of up to 6 MRI visits per patient alongside 13 genomic attributes; it serves as our primary development benchmark (90 training, 19 validation, 19 test patients; context shape [6,64,192,192]). **CFB-GBM** (Moreau et al., 2025) provides an additional glioblastoma cohort with up to 3 longitudinal MRI visits per patient but lacks genomic profiling (93 training, 18 validation, 18 test patients; context shape [3,64,192,192]). Finally, **UCSF-ALPTDG** (Fields et al., 2024) represents a diffuse glioma dataset focused on short-term dynamics across 2 consecutive post-treatment follow-up scans (140 training, 32 validation, 32 test patients; context shape [2,64,96,96]).

To maximize training efficiency, each patient trajectory $(v_1, \ldots, v_T)$ is converted into multiple history–target pairs. For each $1 \leq i < T$, the history is the temporally ordered prefix $(v_1, \ldots, v_i)$, and the target is the immediately subsequent visit v_{i+1}.

Evaluation Metrics. We report MSE (↓), SSIM (↑), and PSNR (↑) to assess image fidelity. To evaluate longitudinal change, let $\Delta = I_1 - I_0$ and $\hat{\Delta} = \hat{I} * 1 - I_0$ denote the observed and predicted residuals. Residual Pattern Alignment measures their spatial and directional agreement independently of overall magnitude:

$$R_{\text{Pattern}} = \frac{\langle \hat{\Delta}, \Delta \rangle}{|\hat{\Delta}|_2 |\Delta|_2}. \tag{17}$$

Scores range from -1 (opposite change) to 1 (perfect alignment). Complementarily, the change magnitude ratio $|\hat{\Delta}|_2/|\Delta|_2$ quantifies the amount of predicted change: 0 corresponds to copying the previous scan, 1 indicates matched magnitude, and values below or above 1 indicate underestimation or overestimation, respectively.

4 Results and Discussion

OncoFlow uses soft source conditioning on the previous visit. **+CC** additionally conditions generation on clinical covariates: treatment information and, for MU-Glioma only, 13 genomic variables. **+HD** adds horizon decay. All baseline models were retrained at the same resolution and evaluated on identical dataset splits. For evaluation, all reported results are computed on the test split using the full sequence context (predicting the final visit as target without window sampling). The *Copy Baseline* returns the previous visit and therefore scores 0 on R_{Pattern} and CMRS by construction. Its SSIM (80.01 on MU-Glioma, 91.76 on CFB-GBM, 95.54 on UCSF-ALPTDG) also serves as a proxy for how much each cohort changes between visits.

Forecasting change without hallucinating anatomy. Figure 2 shows the core difficulty of the task: a forecast must capture localized tumor evolution while leaving stable anatomy unchanged. OncoFlow-Noise+CC, which is initialized from pure Gaussian noise and uses no segmentation of the previous scan, still recovers the location and extent of progression. This indicates that the model forecasts future disease states from its clinically conditioned velocity field rather than by copying the input or relying on the lesion mask. Its main failure mode is spurious residuals in distant healthy tissue. The soft lesion prior and horizon decay suppress these residuals, so the role of the mask is to preserve stable anatomy rather than to enable change forecasting. The cost is a slightly attenuated change within the lesion, but the resulting change map remains the closest to the ground truth. In contrast, CRONOS and ImageFlowNet predict negligible change.

Quantitatively, because most voxels are stable between visits, image-similarity metrics are dominated by unchanged anatomy. On MU-Glioma, the cohort with the most pronounced progression, the Copy Baseline outperforms all competing baselines on SSIM and PSNR; only OncoFlow variants match or exceed it. Even without clinical conditioning or horizon decay, the base OncoFlow model outperforms all learned baselines on every metric across all three cohorts. The largest margins are in change fidelity: relative to the strongest baseline, R_{Pattern} rises from 0.20 to 0.27 on MU-Glioma and from 0.31 to 0.40 on CFB-GBM. The architecture and the spatially varying source therefore account for a substantial part of the improvement. With both components added, the full model doubles R_{Pattern} on MU-Glioma relative to the strongest baseline (0.40 vs. 0.20, ImageFlowNet). On CFB-GBM, it is best on all four metrics, with large change-fidelity margins (R_{Pattern} 0.50 vs. 0.31). Since UCSF-ALPTDG contains mostly stable diffuse gliomas (Fig. 3). OncoFlow

still outperforms all learned baselines, but copying the previous scan remains strongest on image similarity, as any method that re-synthesizes the volume incurs a small penalty when little changes.

Table 1: Quantitative evaluation across datasets. All models were retrained at identical resolutions and evaluated on full test split contexts. SADM is trained with OncoFlow's flow-matching objective. *SSIM values are scaled by 10^2. ± denotes standard deviation across patients.

		Visual Metrics		Residual Change Fidelity	
Dataset	**Method**	**SSIM*** ↑	**PSNR** ↑	$R_{\textbf{Pattern}}$ ↑	**CMRS**
MU-Glioma	Copy Baseline	80.01 ± 2.43	20.19 ± 0.83	0	0
	SADM	73.49 ± 2.85	17.14 ± 0.98	0.16 ± 0.08	0.38 ± 0.18
	ImageFlowNet	79.20 ± 1.47	19.85 ± 0.31	0.20 ± 0.04	0.44 ± 0.14
	CRONOS	76.23 ± 1.65	19.73 ± 0.44	0.16 ± 0.02	0.49 ± 0.17
	OncoFlow	81.20 ± 1.13	20.19 ± 0.54	0.27 ± 0.03	0.58 ± 0.10
	OncoFlow+CC	81.70 ± 1.84	20.31 ± 0.81	0.34 ± 0.02	0.75 ± 0.16
	OncoFlow+CC+HD	**81.90 ± 1.57**	**21.10 ± 0.70**	**0.40 ± 0.04**	**0.80 ± 0.18**
CFB-GBM	Copy Baseline	91.76 ± 2.52	25.95 ± 1.56	0	0
	SADM	82.85 ± 2.04	20.10 ± 0.92	0.25 ± 0.07	0.34 ± 0.15
	ImageFlowNet	91.86 ± 1.20	23.91 ± 1.31	0.31 ± 0.06	0.58 ± 0.17
	CRONOS	90.68 ± 2.01	23.19 ± 1.34	0.18 ± 0.09	0.38 ± 0.13
	OncoFlow	91.93 ± 3.27	25.21 ± 1.30	0.40 ± 0.04	0.62 ± 0.10
	OncoFlow+CC	92.86 ± 3.27	26.08 ± 1.28	0.48 ± 0.03	0.89 ± 0.07
	OncoFlow+CC+HD	**93.01 ± 2.11**	**26.33 ± 1.28**	**0.50 ± 0.05**	**0.91 ± 0.08**
UCSF-ALPTDG	Copy Baseline	**95.54 ± 2.60**	**27.44 ± 1.66**	0	0
	SADM	83.17 ± 1.92	22.06 ± 1.58	0.37 ± 0.02	0.65 ± 0.09
	ImageFlowNet	89.66 ± 2.54	25.38 ± 1.34	0.37 ± 0.03	0.76 ± 0.08
	CRONOS	91.00 ± 2.65	25.19 ± 1.74	0.39 ± 0.01	0.82 ± 0.05
	OncoFlow	91.20 ± 2.81	25.65 ± 1.44	0.41 ± 0.02	0.89 ± 0.08
	OncoFlow+CC	93.00 ± 3.76	26.90 ± 0.91	0.42 ± 0.03	0.99 ± 0.09
	OncoFlow+CC+HD	91.31 ± 2.68	25.75 ± 1.16	0.41 ± 0.03	1.09 ± 0.09

Clinical conditioning and horizon decay. Clinical conditioning (+CC) improves every metric on every cohort. The largest gains are in CMRS: 0.58 → 0.75 on MU-Glioma, 0.62 → 0.89 on CFB-GBM, and 0.89 → 0.99 on UCSF-ALPTDG. Horizon decay improves all metrics on MU-Glioma and CFB-GBM. On UCSF-ALPTDG it lowers SSIM and PSNR and moves CMRS past 1, indicating slight over-prediction of change. This matches its mechanism, where relaxing reliance on the previous scan helps when change occurs but introduces small background errors when it does not.

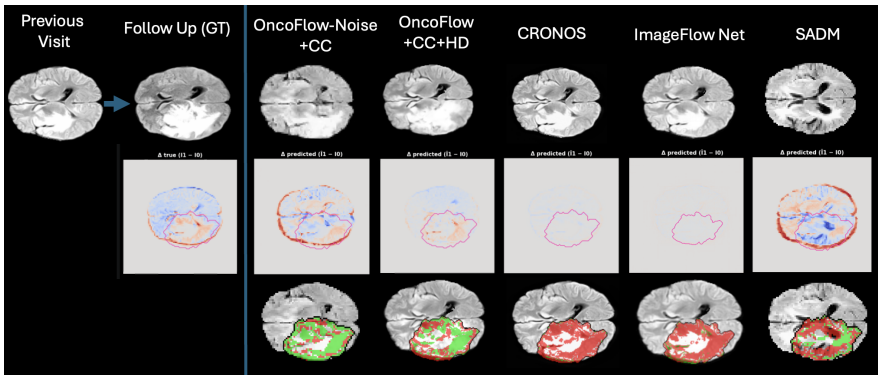

Figure 2: **Qualitative comparison of longitudinal tumor forecasting.** Rows show the predicted follow-up, the residual change from the previous visit, and tumor-region agreement. OncoFlow+CC+HD best balances tumor evolution against preservation of healthy anatomy.

Table 2: **Ablation study on model components.** Evaluation of different conditioning source representations, Horizon Decay, loss formulations, and image spatial resolutions (96×96 vs. 192×192) on reconstruction performance across SSIM (↑) and PSNR (↑).

	Source			Horizon	Clinical	Loss			Resolution		Evaluation Metrics	
#	Noise	Soft	Hard	Decay	Conditioned	Standard	Weighted	Curriculum	96×96	192×192	SSIM ↑	PSNR ↑
1			✓		✓	✓			✓		0.7563	19.47
2			✓		✓		✓		✓		0.7690	19.59
3	✓				✓		✓		✓		0.7740	19.14
4		✓			✓		✓		✓		0.7950	19.61
5		✓			✓	✓			✓		0.8100	20.04
6		✓		✓	✓	✓			✓		0.8110	20.13
7		✓						✓		✓	0.8120	20.19
8		✓			✓			✓		✓	0.8170	20.31
9		✓		✓	✓			✓		✓	**0.8190**	**21.10**

Ablation Study. Table 2 shows that soft dilation consistently outperforms hard dilation and pure-noise initialization under matched loss and resolution, improving all three metrics; SSIM increases to 0.7950, compared with 0.7690 and 0.7740, respectively. For soft dilation, standard loss further improves reconstruction over weighted loss. Horizon decay provides additional gains at both resolutions, increasing PSNR from 20.31 to 21.10 dB in the higher-resolution curriculum configuration. This configuration achieves the highest SSIM of 0.8190 and PSNR.

5 Conclusion

This work presents OncoFlow, a clinically conditioned flow-matching framework for brain MRI forecasting that couples a spatially varying, horizon-dependent source with a velocity network conditioned on longitudinal imaging memory and clinical context. Although this source is constructed from the lesion mask of the previous scan, OncoFlow's forecasting ability does not depend on it: a mask-free variant of OncoFlow initialized from pure Gaussian noise still captures the location and extent of tumor change. The mask therefore acts mainly as an anatomical anchor that suppresses false change, rather than as the source of the forecast.

Across three longitudinal glioma cohorts, OncoFlow outperforms all learned baselines on both image-similarity and change-fidelity metrics, with the largest gains where there is change to forecast. On the predominantly stable UCSF-ALPTDG cohort, copying the previous scan remains strongest on image similarity, and the fixed, clinically motivated horizon-decay schedule slightly over-predicts change, suggesting that the preservation schedule should adapt to the expected magnitude of change rather than to elapsed time alone. More broadly, this exposes how reconstruction metrics can reward forecasts that do not forecast, and indicates that rigorous evaluation of longitudinal generative models requires a copy baseline together with change-aware metrics that disentangle the direction and magnitude of predicted change.

The scope of these conclusions is shaped by the data available. Public longitudinal brain-tumor cohorts with aligned imaging and treatment remain scarce and small, which limits test-set size. Natural next steps are to extend OncoFlow to multi-sequence MRI evaluated with tumor-level clinical endpoints, and study treatment-conditioned forecasts in settings that support causal interpretation.

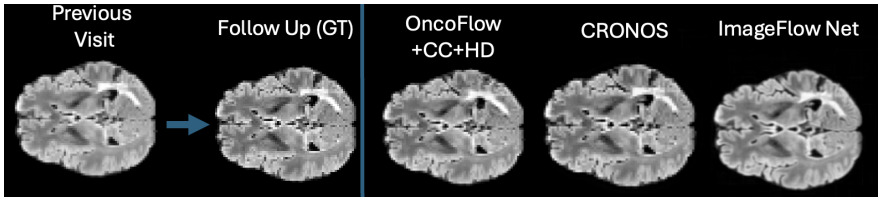

Figure 3: **Conservative forecasts on UCSF-ALPTDG.** In this mostly stable cohort, predictions that stay close to the previous visit achieve high reconstruction scores despite modeling little change. This narrows the visual-metric gap to OncoFlow.

AI use statement

In this work, we used generative AI tools for writing assistance. We have not used generative AI tools for research ideation or conceptualizing the methodological framework, and data generation and experimental execution are not applicable to this work. Additionally, we used generative AI tools for code refactoring, generating plot visualization scripts, and formatting literature citations. We have reviewed all AI-assisted work. Specifically, all LLM-assisted prose was thoroughly edited by the first author to ensure scientific accuracy, clarity, and original phrasing; all bibliographic entries produced with AI assistance were cross-referenced directly against official publisher metadata to prevent hallucinated references. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

To ensure the full reproducibility of our work, comprehensive implementation details including model architectures and hyperparameters are provided in the Supplementary Material. Furthermore, complete source code, preprocessing scripts, and configuration files required to train and evaluate OncoFlow are available in our anonymous repository at `https://anonymous.4open.science/r/OncoFlow-DE5D/`.

References

Valentin Biller, Niklas Bubeck, Lucas Zimmer, Ayhan Can Erdur, Sandeep Nagar, Anke Meyer-Baese, Daniel Rückert, Benedikt Wiestler, and Jonas Weidner. TumorFlow: Physics-guided longitudinal MRI synthesis of glioblastoma growth. *arXiv preprint arXiv:2603.04058*, 2026.

Nico Albert Disch, Saikat Roy, Constantin Ulrich, Yannick Kirchhoff, Maximilian Rokuss, Robin Peretzke, David Zimmerer, and Klaus Maier-Hein. Cronos: Continuous time reconstruction for 4d medical longitudinal series. *International Conference on Learning Representations*, 2026.

Cheng-Chen Fan, Liang Peng, Tian Wang, Hailiang Yang, Xiao-Hu Zhou, Zi-Lian Ni, Guang Wang, Shuo Chen, Yu-Jie Zhou, and Zeng-Guang Hou. Tr-gan: Multi-session future mri prediction with temporal recurrent generative adversarial network. *IEEE Transactions on Medical Imaging*, 41(8):1925–1937, 2022. doi: 10.1109/TMI.2022.3151118.

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), 2024. doi: 10.1148/ryai.230182.

Haotian Jiang, Kai Yu, Meng Li, Yun Cui, Xiaohui Ren, Chen Yang, Xiaoming Zhao, and Song Lin. Classification of progression patterns in glioblastoma: Analysis of predictive factors and clinical implications. *Frontiers in Oncology*, 10:590648, 2020. ISSN 2234-943X. doi: 10.3389/fonc.2020.590648. URL `https://doi.org/10.3389/fonc.2020.590648`.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *International Conference on Learning Representations*, 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 *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 444–453. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72069-7_42.

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, 2025a. doi: 10.1109/ICASSP49660.2025.10890535.

Qinghui Liu, Elies Fuster-Garcia, Ivar Thokle Hovden, Bradley J. MacIntosh, Edvard Grødem, Petter Brandal, Carles Lopez-Mateu, Donatas Sederevicius, Karoline Skogen, Till Guenter 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, 2025b. doi: 10.1109/TMI.2025.3533038.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11461–11471, 2022.

E. Mahmoud, J. Gass, Y. Dhemesh, et al. MU-Glioma Post: A comprehensive dataset of automated MR multi-sequence segmentation and clinical features. *Scientific Data*, 12(1):1847, 2025. doi: 10.1038/s41597-025-06011-7.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Michael T. Milano, Paul Okunieff, Rosemary S. Donatello, Nimish A. Mohile, Joohee Sul, Kevin A. Walter, and David N. Korones. Patterns and timing of recurrence after temozolomide-based chemoradiation for glioblastoma. *International Journal of Radiation Oncology*Biology*Physics*, 78(4):1147–1155, 2010. ISSN 0360-3016. doi: 10.1016/j.ijrobp.2009.09.018. URL `https://doi.org/10.1016/j.ijrobp.2009.09.018`.

N. Moreau, A. Leclercq, A. Desmonts, Y. Poirier, A. Dubru, L. Guillemette, P. Lecoeur, K. Lemasson, C. Jaudet, C. Brunaud, S. Valable, J. Geffrelot, D. Stefan, T. Leleu, C. Raboutet, L. Le Henaff, A. Batalla, J. Lacroix, R. Rouzier, and A. Corroyer-Dulmont. Pre and post treatment MRI and radiotherapy plans of patients with glioblastoma: The CFB-GBM cohort, 2025. URL `https://doi.org/10.7937/V9PN-2F72`. Dataset.

National Cancer Institute. The economic burden of cancer. In *Cancer Trends Progress Report*. National Institutes of Health, U.S. Department of Health and Human Services, 2024. URL `https://progressreport.cancer.gov/`.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. Enhancing spatiotemporal disease progression models via latent diffusion and prior knowledge. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15002 of *Lecture Notes in Computer Science*, pp. 173–183. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72069-7_17.

Patrick Y. Wen, Susan M. Chang, Martin J. Van den Bent, Michael A. Vogelbaum, David R. Macdonald, Eudocia Q. Lee, Michael Weller, Olivier L. Chinot, David A. Reardon, Jann N. Sarkaria, J. Robert Macdonald, Chantal Belanger, Keith Warren, Lauren E. Abrey, and Timothy F. Cloughesy. Response assessment in neuro-oncology clinical trials. *Journal of Clinical Oncology*, 35(21):2439–2449, 2017. doi: 10.1200/JCO.2017.72.7511.

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 (IPMI)*, pp. 388–400. Springer, 2023. doi: 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 *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.