TUG: TUMOR-GUIDED LONGITUDINAL MRI FORECASTING WITH TRAJECTORY SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting longitudinal MRI requires modeling tumor evolution while preserving anatomical realism, but these objectives are not always aligned. Image-oriented generative models may produce high-fidelity scans while underrepresenting tumor changes, whereas stronger tumor supervision may compromise image quality. We introduce TUG, a tumor-guided forecasting framework that bridges this appearance–evolution gap during training and inference. During training, we introduce an evolution-aware residual modulator that injects tumor structure, follow-up interval, and treatment information into the generative residuals. We further supervise the lesion expressed by one-step clean-image estimates and preserve future-tumor information in intermediate diffusion features. During inference, we introduce tumor-aware candidate scoring to progressively prune low-ranked trajectories and softly weight the completed predictions using terminal tumor evidence. Across three longitudinal glioma MRI cohorts, our approach consistently improves tumor-evolution fidelity while retaining competitive image quality, demonstrating complementary benefits from training and inference time.

1 INTRODUCTION

Longitudinal MRI forecasting aims to generate plausible future scans from a patient's current imaging, together with temporal and treatment information, providing an image-based means to characterize disease course and treatment response Yoon et al. (2023); Puglisi et al. (2025); Liu et al. (2025); Wang et al. (2026). An effective forecast should preserve future anatomy while accurately representing tumor-related changes such as growth and regression. Yet many generative approaches are optimized with image-level objectives and evaluated primarily using whole-image measures such as PSNR and SSIM Yoon et al. (2023); Puglisi et al. (2025); Gourdeau et al. (2022). Because tumor-related changes occupy only a small portion of the brain, a prediction can achieve strong global fidelity while misrepresenting clinically important changes in the tumor-associated region. Recent methods have therefore introduced explicit tumor-related supervision through, for example, joint future-mask prediction or representation alignment Liu et al. (2025); Wang et al. (2026). Joint optimization remains nontrivial: whole-image objectives are dominated by largely stable normal anatomy, whereas tumor-specific objectives concentrate on a small and highly variable region. Strengthening tumor supervision may consequently improve lesion consistency while degrading the anatomical or intensity fidelity of the generated MRI.

Image fidelity and tumor-evolution fidelity are related but not interchangeable. Restricting image-fidelity evaluation to the tumor neighborhood remains insufficient because local voxel-wise similarity primarily reflects appearance agreement and does not explicitly characterize lesion morphology or its evolution. A small boundary displacement, missed growth region, or attenuated tumor contrast may introduce only modest intensity error while substantially altering the lesion inferred from the generated image. More generally, image-based similarity measures provide limited information about localized disease-related changes Gourdeau et al. (2022); Puglisi et al. (2025). This limitation is particularly relevant in stochastic forecasting. Different trajectories can yield distinct local tumor realizations, while averaging repeated predictions—a strategy used to stabilize longitudinal diffusion forecasts Puglisi et al. (2025)—may suppress spatially varying structures and produce a smoother prediction with favorable voxel-wise fidelity. Evaluating longitudinal generation therefore requires considering not only resemblance to the observed future, but also whether the generated image faithfully expresses the corresponding tumor changes.

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

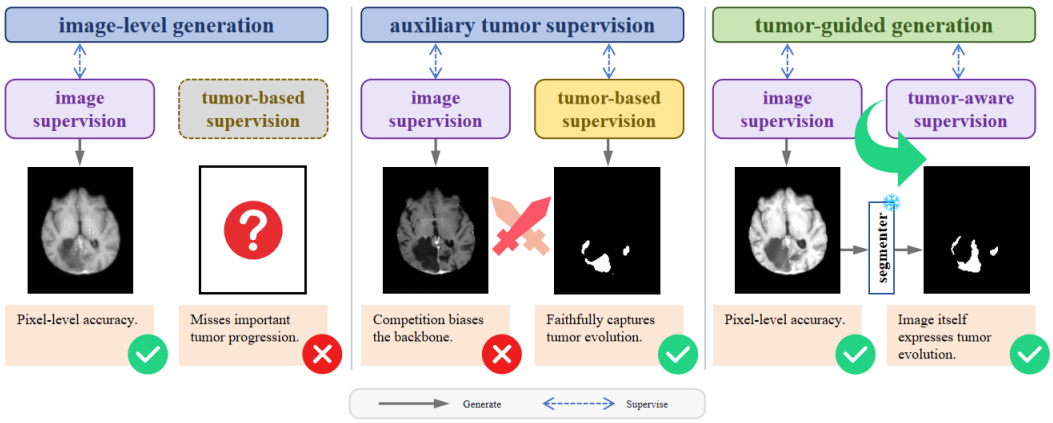

Figure 1: **Motivation for tumor-guided longitudinal MRI forecasting.** Whole-image objectives can yield high voxel-wise fidelity while overlooking tumor evolution, whereas stronger tumor supervision may improve lesion modeling at the expense of image fidelity. We incorporate tumor information during both training and inference to balance these objectives.

A natural way to incorporate tumor information is to predict the future lesion mask as an auxiliary task or introduce tumor structure into the learned representation Liu et al. (2025); Wang et al. (2026). Such strategies strengthen disease-related modeling but do not eliminate the imbalance between whole-image generation and localized tumor supervision. Moreover, future lesion geometry is difficult to predict because tumor appearance and response vary with treatment, imaging interval, and patient-specific disease dynamics Liu et al. (2025); Wen et al. (2023). Rather than prescribing a deterministic future lesion mask before generation, we introduce TUG, which incorporates tumor guidance directly into longitudinal MRI forecasting. A lightweight evolution-aware residual modulator injects current lesion geometry, elapsed time, and treatment information into the conditional generation pathway. Training further employs two complementary objectives: segmentation-based supervision is applied to lesion structure recovered from one-step clean-image estimates, while an alignment objective encourages intermediate generative features to retain information predictive of the future lesion state.

Tumor guidance also informs how stochastic predictions are used at inference. Although trajectories originate from the same patient condition, their eventual tumor fidelity can vary substantially, and we observe that informative differences become progressively identifiable during denoising. We therefore introduce tumor-guided trajectory selection and aggregation. A prospective scorer uses tumor-related evidence to progressively prune low-ranked partially denoised trajectories, while a terminal scorer softly weights the completed candidates for aggregation. In this way, tumor-related information guides not only model learning but also the selection and aggregation of stochastic predictions. Across three longitudinal glioma MRI cohorts, the resulting framework improves tumor-evolution fidelity while retaining competitive overall image quality.

Our contributions are summarized as follows:

- We introduce a tumor-aware training framework that combines lightweight evolution-aware residual modulation with supervision on generated lesion structure and intermediate representations, balancing tumor-evolution fidelity and image fidelity.
- We develop tumor-guided trajectory selection and aggregation, progressively scoring and pruning partially denoised trajectories and softly weighting completed predictions using terminal tumor evidence.
- Experiments on three longitudinal glioma MRI cohorts demonstrate improved fidelity of predicted tumor changes while retaining competitive future-image quality, with complementary gains from training-time guidance and inference-time selection and aggregation.

2 Related Work

2.1 Longitudinal Medical Image Forecasting

Longitudinal medical image forecasting has progressed from explicit disease-growth modeling toward data-driven prediction of patient-specific future anatomy. Earlier work on brain tumors incorporated biophysical growth priors or learned longitudinal lesion dynamics, while GP-GAN introduced stacked 3D conditional GANs with segmented feature maps and joint intensity- and Dice-based supervision Elazab et al. (2020). TaDiff jointly predicts treatment-conditioned multiparametric MRI and future tumor masks at arbitrary target times Liu et al. (2025), whereas Brain Latent Progression (BrLP) uses 3D latent diffusion with subject metadata and disease-progression priors and aggregates repeated stochastic predictions through Latent Average Stabilization Puglisi et al. (2025).

Recent models incorporate richer temporal and clinical context. CRONOS supports continuous-time 3D forecasting Disch et al. (2026); 3D-GlioPREDICT conditions volumetric latent diffusion on follow-up time and radiotherapy dose Belkacemi et al. (2026); Brain-WM jointly models future MRI and treatment planning Wang et al. (2026); and TumorFlow combines biophysical dynamics with generative MRI synthesis Biller et al. (2026). Broader medical world-model work connects clinical state, dynamics, and intervention modeling Liu et al. (2026a), alongside recent consistency and implicit-representation approaches Liu et al. (2026b; 2023); Li et al. (2025a;b) and multimodal or agentic medical AI systems Shi et al. (2025); Li et al. (2026). Our work instead uses tumor information to guide both future-MRI generation and stochastic inference.

2.2 Tumor-Guided Generative Supervision

Structural supervision is widely used to emphasize clinically relevant anatomy during image generation. GP-GAN incorporates segmented features and Dice supervision, TaDiff jointly predicts MRI and tumor segmentation, Brain-WM aligns intermediate representations using longitudinal tumor masks, and MedSegFactory couples image and segmentation diffusion streams Elazab et al. (2020); Liu et al. (2025); Wang et al. (2026); Mao et al. (2025).

Related controllable-diffusion methods use discriminative feedback on denoised estimates or decode semantic structure from intermediate diffusion features for guidance and consistency, as in ControlNet++, Ctrl-U, Readout Guidance, and InnerControl Li et al. (2024); Zhang et al. (2025a); Luo et al. (2024); Konovalova et al. (2025). REPA further demonstrates the value of supervising intermediate generative representations Yu et al. (2025b).

Our setting differs because the desired future tumor is unavailable at deployment. We use it only as training supervision: current tumor geometry is injected through lightweight residual modulation, image-space supervision constrains the lesion expressed by clean-image estimates, and representation-space supervision preserves information predictive of the future tumor state.

2.3 Inference-Time Trajectory Refinement for Diffusion Models

Stochastic diffusion naturally produces multiple predictions from the same condition. BrLP exploits this variability through repeated sampling and averaging Puglisi et al. (2025), while broader inference-time methods use learned verifiers or search to favor higher-quality trajectories Ma et al. (2025). Noise Trajectory Search and Diffusion Tree Sampling extend selection beyond the initial seed to decisions made throughout denoising Ramesh & Mardani (2025); Jain et al. (2025). More recent methods estimate trajectory quality before generation is complete: TTSnap aligns intermediate scores with terminal rewards, while Progressive Seed Pruning progressively removes less promising candidates using intermediate denoised estimates Yu et al. (2025a); Guimaraes & Perona (2026).

Longitudinal tumor forecasting differs because true future-tumor agreement cannot be measured at inference time. We therefore learn trajectory scorers using only deployment-available evidence. A prospective scorer progressively prunes partially denoised trajectories, while a terminal scorer softly reweights the completed predictions. Both stages use tumor-related evidence, extending tumor guidance from model training to stochastic forecast refinement.

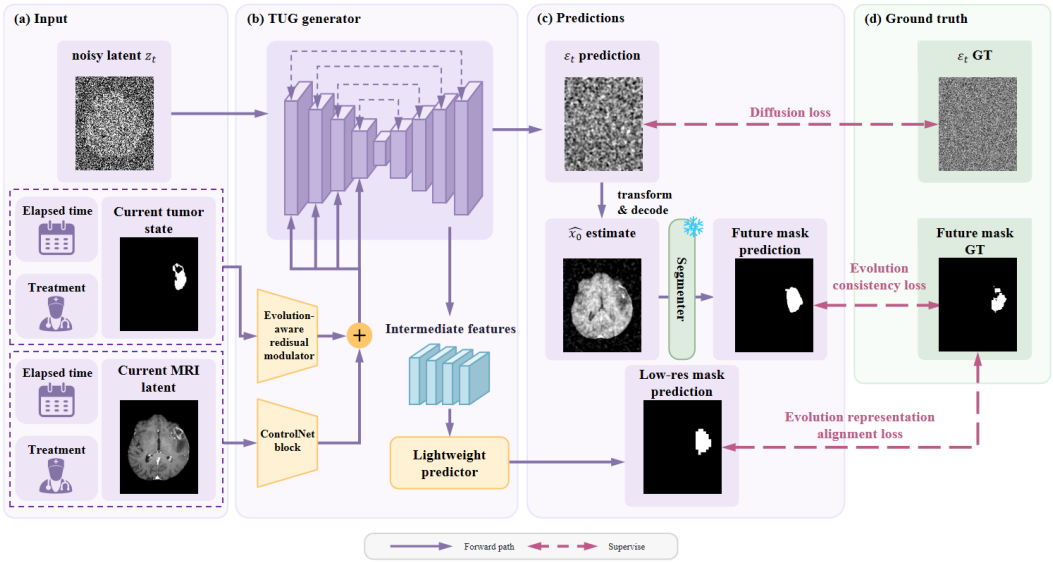

Figure 2: **Training overview of TUG.** (a) Latent-space diffusion inputs and patient-specific conditions. (b) ERM modulates the conditional residuals injected into the diffusion UNet. (c) The predicted noise supports image-space evolution consistency and representation-space evolution alignment in addition to the diffusion objective. (d) Sampled noise and the observed future mask provide the training targets. Solid arrows indicate forward computation and dashed magenta arrows indicate supervision.

3 Tumor-Guided Longitudinal MRI Forecasting

3.1 Overview and Problem Formulation

Given a current MRI X_s, tumor mask M_s, elapsed time Δt, and treatment information c, we forecast a plausible future MRI X_f. The future image and tumor mask M_f are available only during training; inference uses only the current examination and generated trajectories. TUG builds on a conditional 3D latent diffusion model and introduces tumor guidance at three points: conditional generation, tumor-specific training supervision, and stochastic trajectory refinement. Figure 2 summarizes the training pathway.

During training, an evolution-aware residual modulator injects current tumor structure and clinical context into the conditional residual pathway. Two complementary objectives then constrain the tumor expressed by one-step clean-image estimates and preserve future-tumor information in intermediate diffusion features. At inference, a prospective scorer prunes less promising partial trajectories, after which a terminal scorer softly weights the completed survivors.

Importantly, tumor guidance constrains the image-generation process rather than prescribing a deterministic future lesion mask. The current lesion informs conditional generation, whereas the observed future lesion is used only for training-time supervision. The final forecast therefore remains a stochastic MRI prediction whose tumor state is determined by what is expressed in the generated image.

3.2 Tumor-Guided Generator

3.2.1 Conditional Latent Diffusion Backbone

Following BrLP Puglisi et al. (2025), we use a conditional 3D latent diffusion backbone with a frozen image autoencoder and diffusion UNet together with a trainable ControlNet pathway Zhang et al. (2023). The model follows the standard noise-prediction objective, and DDIM sampling Song et al. (2021) produces multiple stochastic futures. We initialize from the pretrained forecasting

model and augment only its conditional pathway; the full backbone formulation is given in Appendix B.

3.2.2 EVOLUTION-AWARE RESIDUAL MODULATOR

We introduce an **Evolution-Aware Residual Modulator (ERM)** between selected ControlNet outputs and the diffusion UNet. A lightweight spatial encoder extracts multi-scale tumor features from the current examination, which are combined with an embedding of diffusion timestep t, follow-up interval Δt, and treatment c. At each modulated scale l, ERM predicts spatially varying affine parameters γ_l and β_l and transforms the ControlNet residual R_l as

$$\widetilde{R}_l = (1 + \gamma_l) \odot R_l + \beta_l. \tag{1}$$

The modulation is applied sparsely and initialized to preserve the original ControlNet pathway, allowing tumor and evolution-related context to guide the conditional residuals without modifying the pretrained diffusion backbone. Spatial inputs, multi-scale construction, and initialization are detailed in Appendix B.

3.3 TUMOR-SPECIFIC TRAINING SUPERVISION

Conditioning alone does not ensure that the synthesized MRI expresses the observed lesion evolution. We therefore use the future tumor mask M_f only as training supervision at two complementary levels. For predicted lesion logits ℓ, both objectives share

$$\mathcal{L}_{\mathrm{tum}}(\ell, M_f; M_s) = \lambda_{\mathrm{D}} \mathcal{L}_{\mathrm{Dice}}(\sigma(\ell), M_f) + \lambda_{\mathrm{B}} \mathcal{L}_{\mathrm{WBCE}}(\ell, M_f; M_s), \tag{2}$$

where the weighted BCE emphasizes tumor-related voxels over background. Further loss details are provided in Appendix B.

3.3.1 IMAGE-SPACE EVOLUTION CONSISTENCY

Given a noisy latent z_t, we recover a one-step clean estimate

$$\hat{z}_0^{(t)} = \frac{z_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(z_t, t, \mathcal{C})}{\sqrt{\bar{\alpha}_t}}, \qquad \hat{X}_0^{(t)} = \mathcal{D}(\hat{z}_0^{(t)}). \tag{3}$$

A separately trained, frozen tumor segmenter S extracts the lesion expressed by this image. The **Evolution Consistency Loss (ECL)** is

$$\mathcal{L}_{\mathrm{ECL}} = \mathcal{L}_{\mathrm{tum}} \left(S(\hat{X}_0^{(t)}), M_f; M_s \right). \tag{4}$$

ECL therefore supervises tumor structure through the generated image rather than through an independent future-mask branch, and is applied only after sufficient denoising for image-space structure to be reliable.

3.3.2 REPRESENTATION-SPACE EVOLUTION ALIGNMENT

To provide tumor guidance over a broader portion of denoising, we additionally use a frozen lightweight readout P_{ψ^*} that predicts low-resolution future-lesion logits from multi-level UNet features $\mathcal{H}_t$ and the timestep embedding e_t. This yields the **Evolution Representation Alignment (ERA)** loss

$$\mathcal{L}_{\mathrm{ERA}} = \mathcal{L}_{\mathrm{tum}} \left(P_{\psi^*}(\mathcal{H}_t, e_t), M_f; M_s \right). \tag{5}$$

Unlike ECL, ERA does not require a fully formed clean image; it preserves the ability of intermediate generative features to predict the future tumor state. The readout is pretrained and then kept fixed during tumor-guided fine-tuning.

3.3.3 OVERALL TRAINING OBJECTIVE

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{ECL}} \mathbf{1}[t \in \mathcal{T}_{\mathrm{ECL}}] \mathcal{L}_{\mathrm{ECL}} + \lambda_{\mathrm{ERA}} \mathbf{1}[t \in \mathcal{T}_{\mathrm{ERA}}] \mathcal{L}_{\mathrm{ERA}}, \tag{6}$$

where the two tumor losses are activated over their respective timestep ranges. The autoencoder, diffusion UNet, segmenter, and evolution readout remain frozen, while the conditional pathway and ERM are optimized. ECL directly constrains image-expressed tumor structure, whereas ERA supplies complementary representation-level guidance.

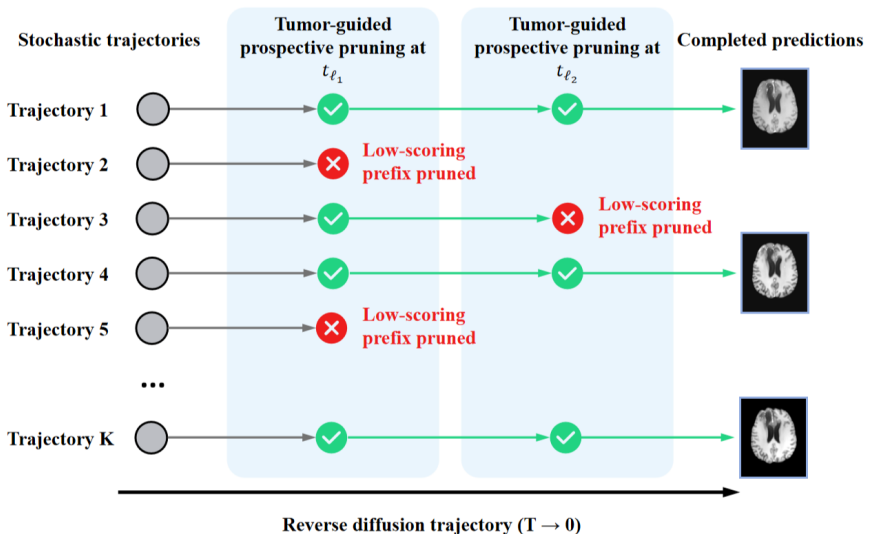

Figure 3: **Tumor-guided prospective trajectory pruning.** At selected denoising steps, low-scoring trajectory prefixes are pruned so that the remaining computation is concentrated on promising candidates.

3.4 Tumor-Guided Stochastic Trajectory Refinement

Stochastic sampling produces multiple plausible futures with different tumor fidelity. Rather than averaging them uniformly, we first prune low-ranked partial trajectories and then softly aggregate the completed survivors.

The two stages address different aspects of stochastic inference: prospective pruning reallocates a fixed sampling budget before generation is complete, whereas soft terminal weighting retains information from multiple completed hypotheses instead of making a single hard selection.

3.4.1 Prospective Trajectory Pruning

Because future-tumor agreement is unavailable at deployment, we train lightweight within-patient rankers whose inputs use only the current examination and generated trajectories. At a pruning timestep t_ℓ, a one-step clean-image estimate and its frozen-segmenter lesion hypothesis are used to construct tumor-, change-, population-, and condition-related descriptors. The scorer ranks the surviving prefixes and retains the top K_ℓ candidates; population-relative descriptors are recomputed after each pruning operation. Training preferences are derived from terminal Dice against the observed future mask, but the future image and mask are never scorer inputs at inference. Descriptor definitions, ranking loss, and the explicit selection operator are provided in Appendix B.

3.4.2 Terminal Soft Weighting

After denoising, a terminal scorer evaluates each survivor using image- and lesion-space evidence together with summaries from the frozen evolution readout. Standardized terminal scores are converted into soft aggregation weights,

$$\bar{s}_k = \frac{s_k^{\text{term}} - \mu_s}{\sigma_s}, \qquad \pi_k = \frac{\exp(\bar{s}_k/\tau)}{\sum_{j\in\mathcal{K}} \exp(\bar{s}_j/\tau)}, \qquad \hat{X}_f = \sum_{k\in\mathcal{K}} \pi_k \hat{X}_f^{(k)}. \tag{7}$$

Lower τ concentrates weight on higher-scoring predictions, whereas higher τ approaches uniform averaging. Prospective pruning therefore allocates computation among partial trajectories, while terminal weighting determines the contribution of completed predictions.

Table 1: Comparison with longitudinal MRI forecasting baselines on the held-out validation set. Higher is better for all metrics.

Model	PSNR ↑	SSIM ↑	Case-avg. Dice ↑	Pooled Dice ↑
SADM	15.51	0.1958	0.0606	0.0766
LDM-GBM	23.94	0.8560	0.4212	0.5298
BrLP	**26.06**	**0.9003**	0.3628	0.4850
TaDiff-3D	19.56	0.5247	<u>0.5155</u>	<u>0.5744</u>
TUG (Ours)	<u>25.15</u>	<u>0.8868</u>	**0.5217**	**0.5873**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and preprocessing. We evaluate on three longitudinal glioma MRI cohorts: LUMIERE Suter et al. (2022), MU-Glioma-Post (MU) Mahmoud et al. (2025), and UCSF-ALPTDG (UCSF) Fields et al. (2024). Temporally adjacent examinations are paired as current and future observations, with elapsed time and treatment information used as additional conditions. Patient-level splitting yields 949 training pairs and 228 held-out validation pairs. All scans are registered to a common spatial reference, resampled to $1.5\,\text{mm}$ isotropic resolution, cropped or padded to $120 \times 144 \times 120$, and intensity-normalized to $[0, 1]$. Dataset composition, preprocessing, and cohort-specific lesion definitions are detailed in Appendix A.

Implementation and inference. All diffusion-based methods use 50 sampling steps. Our inference starts from 24 stochastic trajectories, retains 20 at $t = 400$ and 8 at $t = 200$, and aggregates the completed survivors using terminal soft weighting with $\tau = 1.0$. This schedule has the same nominal denoiser budget as completing 20 independent trajectories; diffusion baselines are therefore evaluated with Mean20 aggregation. Training, trajectory-scorer development, and budget accounting are provided in Appendix B.

Evaluation metrics. We report full-volume PSNR and SSIM for future-image fidelity. Tumor fidelity is evaluated by applying the same fixed nnU-Net-based tumor segmenter used for ECL supervision to each generated MRI and comparing the resulting lesion mask with the dataset-provided future annotation. No method-specific auxiliary mask output is used for evaluation. We report case-averaged Dice, which weights cases equally, and pooled Dice, which aggregates TP, FP, and FN over the cohort and provides a complementary cohort-level measure of lesion overlap.

4.2 COMPARISON WITH LONGITUDINAL FORECASTING BASELINES

We compare against SADM (Yoon et al., 2023), LDM-GBM (Leclercq et al., 2025), BrLP (Puglisi et al., 2025), and TaDiff (Liu et al., 2025); our volumetric adaptation of TaDiff is denoted TaDiff-3D. All methods use the same preprocessing and evaluation protocol and are compared under a matched nominal denoising budget. Task-specific adaptations of SADM and TaDiff-3D are described in Appendix C.

Table 1 shows distinct behavior between future-image and tumor fidelity. BrLP achieves the highest PSNR and SSIM but substantially lower tumor fidelity; relative to BrLP, TUG increases case-averaged Dice from 0.3628 to 0.5217 and pooled Dice from 0.4850 to 0.5873, while retaining $25.1532\,\text{dB}$ PSNR and 0.8868 SSIM.

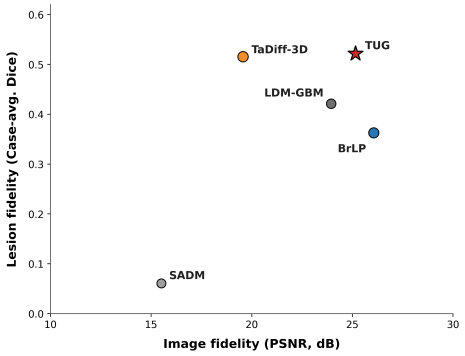

Figure 4: **Image–lesion fidelity trade-off.** TUG combines near-BrLP image fidelity with the strongest case-averaged Dice.

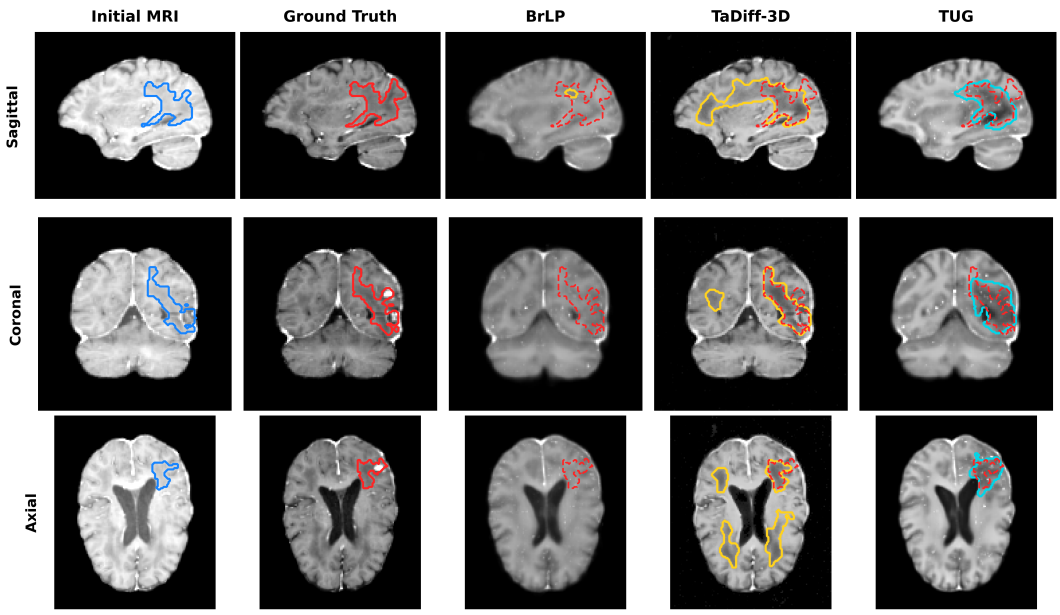

Figure 5: **Qualitative comparison of longitudinal MRI forecasts.** Sagittal, coronal, and axial views are shown. Blue and red contours indicate the initial and ground-truth future tumor regions, respectively; dashed red contours mark the ground truth on predicted images, while yellow and cyan contours show tumor regions recovered from baseline and our predictions. TUG captures the future tumor extent while preserving realistic image appearance.

TaDiff-3D shows the complementary behavior,
achieving strong tumor fidelity but considerably lower image similarity. TUG slightly improves both Dice measures over TaDiff-3D (0.5217 vs. 0.5155 case-averaged and 0.5873 vs. 0.5744 pooled), while improving PSNR by 5.60 dB and SSIM by 0.3621. These results illustrate the appearance–evolution gap motivating our framework: strong global image similarity does not necessarily imply faithful tumor evolution.

Figure 5 complements the quantitative results. In the shown case, BrLP preserves plausible global anatomy but underrepresents parts of the future lesion extent, whereas TaDiff-3D predicts a broader lesion with larger appearance deviations. Our prediction follows the future contour more closely across the three views while retaining the surrounding anatomy, illustrating the need to assess image and lesion fidelity jointly. Because all diffusion-based methods are evaluated under the same nominal denoising budget, these differences cannot be attributed simply to additional sampling computation. In particular, our refinement redistributes the budget across stochastic trajectories rather than increasing the total number of denoiser evaluations. The comparison also highlights why image fidelity and tumor fidelity should be reported jointly: methods with similar or higher global similarity can differ substantially in the lesion structure expressed by the generated MRI, while strong lesion agreement alone does not guarantee realistic whole-image appearance.

4.3 Ablation Studies

We separately ablate tumor-guided training and stochastic trajectory refinement. Training ablations use the same ERM-equipped generator and Mean20 inference, while inference ablations fix the complete ERM+ECL+ERA training configuration.

Tumor-Guided Training As shown in Table 2(a), ECL increases case-averaged Dice from 0.3746 to 0.4549 and pooled Dice from 0.4901 to 0.5362, at the cost of 0.73 dB PSNR and 0.0113 SSIM. Adding ERA further improves the two Dice metrics by 0.0322 and 0.0301, while recovering 0.30 dB PSNR and nearly all of the SSIM reduction introduced by ECL. Overall, ECL provides direct image-

Table 2: Ablation of **(a)** tumor-guided training and **(b)** stochastic trajectory refinement.

(a) Tumor-guided training

Configuration	PSNR ↑	SSIM ↑	Case-avg. Dice ↑	Pooled Dice ↑
No extra loss	**26.05**	**0.8889**	0.3746	0.4901
+ ECL	25.32	0.8776	0.4549	0.5362
+ ECL + ERA	25.61	0.8882	**0.4871**	**0.5663**

(b) Tumor-guided inference

Configuration	PSNR ↑	SSIM ↑	Case-avg. Dice ↑	Pooled Dice ↑
Mean20	**25.61**	0.8882	0.4871	0.5663
Pruning + Mean8	25.39	**0.8903**	0.5089	0.5802
Pruning + Weighting	25.15	0.8868	**0.5217**	**0.5873**

space tumor supervision, while ERA supplies complementary representation-level guidance that further improves tumor fidelity while partially recovering image quality.

Tumor-Guided Stochastic Inference Table 2(b) shows progressive gains from the two inference stages. Prospective pruning increases case-averaged and pooled Dice to 0.5089 and 0.5802, respectively, and terminal weighting further improves them to 0.5217 and 0.5873. Relative to Mean20, the complete refinement strategy gains 0.0346 case-averaged Dice and 0.0210 pooled Dice, with a 0.46 dB PSNR reduction and only a small SSIM change (0.8882 to 0.8868). The two stages are therefore complementary: prospective scoring allocates computation among partial trajectories, whereas terminal scoring differentiates their contributions during final aggregation. Notably, pruning alone already improves both Dice metrics while slightly increasing SSIM relative to Mean20; terminal weighting then provides the remaining tumor-fidelity gain with a modest additional image-quality trade-off.

Taken together, the two ablations indicate complementary effects at different stages of forecasting. Tumor-guided training improves the quality of individual stochastic predictions before any trajectory selection is applied, whereas inference-time refinement further exploits variation among those predictions without modifying the generator. The gains therefore arise from both improving the underlying forecast distribution and making better use of its stochastic samples, rather than from either training supervision or inference-time selection alone.

5 Conclusion and Limitations

We presented TUG, a tumor-guided framework for longitudinal glioma MRI forecasting, which incorporates tumor information into both training and stochastic inference. Evolution-aware residual modulation, image- and representation-space supervision, and trajectory selection and aggregation jointly improve the tumor-evolution fidelity of generated future MRIs while retaining competitive image quality across three longitudinal glioma cohorts. Importantly, training-time guidance and inference-time refinement act at different stages of the forecasting process: the former improves the tumor structure expressed by individual stochastic predictions, while the latter makes better use of the resulting trajectory population under a fixed sampling budget. These results support treating future-image appearance and tumor evolution as complementary objectives, and suggest that disease-specific information can guide not only how generative forecasts are learned, but also how their stochastic predictions are used.

Our study is limited to pairwise T1C forecasting with harmonized treatment conditions, leaving multiple historical examinations, additional modalities, and more detailed clinical variables for future work. Evaluation also relies on a fixed tumor segmenter, and some baselines require adaptation to the volumetric pairwise setting. Independent segmenters, expert assessment, and broader native comparisons would further strengthen the evaluation. Overall, our results demonstrate a practical route to improving disease-related fidelity without abandoning future-image realism.

AI USE STATEMENT

Generative AI tools assisted with literature review, methodological discussions, code development and debugging, result analysis, and the preparation and editing of the manuscript and figures. All AI-assisted outputs, including references, implementations, experimental results, and scientific claims, were reviewed and verified by the authors. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

This study uses established de-identified longitudinal glioma MRI datasets and involves no new human-subject data collection. All datasets were used in accordance with their original licenses and access conditions. The proposed method is intended for research on medical image forecasting and has not been validated for clinical decision-making. Its predictions may inherit biases from the source cohorts, imaging protocols, annotations, and tumor definitions; prospective clinical validation would therefore be required before any practical deployment.

REPRODUCIBILITY STATEMENT

The main paper describes the model architecture, training objectives, inference procedure, datasets, and evaluation protocol. Additional implementation details, including preprocessing, optimization settings, trajectory-scoring features, hyperparameters, and supplementary analyses, are provided in the Appendix. Anonymized code and configuration files will be released to support reproduction of the reported experiments.

REFERENCES

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. *arXiv preprint arXiv:2606.05113*, 2026.

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 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. In *International Conference on Learning Representations*, 2026.

Ahmed Elazab, Changmiao Wang, Syed Jamal Safdar Gardezi, Hongmin Bai, Qingmao Hu, Tianfu Wang, Chunqi Chang, and Baiying Lei. GP-GAN: Brain tumor growth prediction using stacked 3d generative adversarial networks from longitudinal mr images. *Neural Networks*, 132:321–332, 2020.

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.

Daniel Gourdeau, Simon Duchesne, and Louis Archambault. On the proper use of structural similarity for the robust evaluation of medical image synthesis models. *Medical Physics*, 49(4): 2462–2474, 2022.

Rogerio Guimaraes and Pietro Perona. Inference-time scaling of diffusion models via progressive seed pruning. *arXiv preprint arXiv:2607.21591*, 2026.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Vineet Jain, Kusha Sareen, Mohammad Pedramfar, and Siamak Ravanbakhsh. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Nina Konovalova, Maxim Nikolaev, Andrey Kuznetsov, and Aibek Alanov. Heeding the inner voice: Aligning ControlNet training via intermediate features feedback. *arXiv preprint arXiv:2507.02321*, 2025.

Alexandre G. Leclercq, Sébastien Bougleux, Noémie N. Moreau, Alexis Desmonts, Romain Hérault, and Aurélien Corroyer-Dulmont. Pre to post-treatment glioblastoma mri prediction using a latent diffusion model. In *MICCAI Workshop on Deep Generative Models*, pp. 182–191. Springer, 2025. doi: 10.1007/978-3-032-05472-2_18.

Mengxuan Li, Ke Liu, Hongyang Chen, Jiajun Bu, Hongwei Wang, and Haishuai Wang. TSINR: Capturing temporal continuity via implicit neural representations for time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 671–682, 2025a. doi: 10.1145/3690624.3709266.

Mengxuan Li, Ke Liu, Jialong Guo, Jiajun Bu, Hongwei Wang, and Haishuai Wang. ImputeINR: Time series imputation via implicit neural representations for disease diagnosis with missing data. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 9241–9249, 2025b. doi: 10.24963/ijcai.2025/1027.

Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. ControlNet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pp. 129–147. Springer, 2024.

Shiyu Li, Ziqi Yan, Zhihao Wu, Jielong Lu, Weiran Liao, Jiajun Yu, Genjie Li, Zeyu Chu, Jiajun Bu, and Haishuai Wang. DeepBD: A grounded agentic workflow for variant prioritization and diagnosis of genetic birth defects. *arXiv preprint arXiv:2606.24779*, 2026.

Ke Liu, Feng Liu, Haishuai Wang, Ning Ma, Jiajun Bu, and Bo Han. Partition speeds up learning implicit neural representations based on exponential-increase hypothesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5474–5483, 2023.

Ke Liu, Mengxuan Li, Yanyi Bao, Tianyun Zhang, Chong Chu, Jiajun Bu, and Haishuai Wang. Medical world models: Representing medical states, modelling clinical dynamics and guiding intervention policies. *arXiv preprint arXiv:2606.16721*, 2026a.

Ke Liu, Feng Liu, Jingjun Gu, Shuyi Zhang, Zhihua Wang, Jiajun Bu, Bo Han, and Haishuai Wang. One-step image-function generation via consistency training. *IEEE Transactions on Multimedia*, 2026b. doi: 10.1109/TMM.2026.3676205.

Qinghui Liu, Elies Fuster-Garcia, Ivar Thokle Hovden, Bradley J. MacIntosh, Edvard O. S. Grødem, Petter Brandal, Carles Lopez-Mateu, Donatas Sederevicius, 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, 2025.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8217–8227, 2024.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2523–2534, 2025.

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*, 12:1847, 2025. doi: 10.1038/s41597-025-06011-7.

Jiawei Mao, Yuhan Wang, Yucheng Tang, Daguang Xu, Kang Wang, Yang Yang, Zongwei Zhou, and Yuyin Zhou. MedSegFactory: Text-guided generation of medical image-mask pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21525–21535, 2025.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. Brain latent progression: Individual-based spatiotemporal disease progression on 3d brain mris via latent diffusion. *Medical Image Analysis*, 106:103734, 2025.

Vignav Ramesh and Morteza Mardani. Test-time scaling of diffusion models via noise trajectory search. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Keyue Shi, Qianqian Shen, Zhaoming Ye, Liangjun Jiang, Jiajun Bu, and Haishuai Wang. LUMOS: A lumbar multimodal osteoporosis screening dataset with x-ray and ct images. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13250–13257, 2025. doi: 10.1145/3746027.3758282.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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, 2022. doi: 10.1038/s41597-022-01881-7.

Chenhui Wang, Boyun Zheng, Liuxin Bao, Zhihao Peng, Peter Y. M. Woo, Hongming Shan, and Yixuan Yuan. Brain-WM: Brain glioblastoma world model. *arXiv preprint arXiv:2603.07562*, 2026.

Patrick Y. Wen, Martin van den Bent, Gilbert Youssef, Timothy F. Cloughesy, Benjamin M. Ellingson, Michael Weller, Evanthia Galanis, Daniel P. Barboriak, John de Groot, Mark R. Gilbert, Raymond Huang, Andrew B. Lassman, Minesh Mehta, Annette M. Molinaro, Matthias Preusser, Rifaquat Rahman, Lalitha K. Shankar, Roger Stupp, Javier E. Villanueva-Meyer, David A. Reardon, Michael A. Vogelbaum, and Susan M. Chang. RANO 2.0: Update to the response assessment in neuro-oncology criteria for high- and low-grade gliomas in adults. *Journal of Clinical Oncology*, 41:5187–5199, 2023.

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, pp. 388–400. Springer, 2023.

Qingtao Yu, Changlin Song, Minghao Sun, Zhengyang Yu, Vinay Kumar Verma, Soumya Roy, Sumit Negi, Hongdong Li, and Dylan Campbell. TTSnap: Test-time scaling of diffusion models via noise-aware pruning. *arXiv preprint arXiv:2511.22242*, 2025a.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025b.

Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. Ctrl-U: Robust conditional image generation via uncertainty-aware reward modeling. In *International Conference on Learning Representations*, 2025a.

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*, pp. 3836–3847, 2023.

Xinru Zhang, Ni Ou, Berke Doga Basaran, Marco Visentin, Mengyun Qiao, Renyang Gu, Paul M. Matthews, Yaou Liu, Chuyang Ye, and Wenjia Bai. A foundation model for lesion segmentation on brain mri with mixture of modality experts. *IEEE Transactions on Medical Imaging*, 44(6): 2594–2604, 2025b. doi: 10.1109/TMI.2025.3540809.

A Dataset and Preprocessing Details

Pair construction and splits. We construct forecasting instances from temporally adjacent examinations of the same patient, using the earlier examination as the current observation and the later examination as the future target. Splits are performed at the patient level independently within each cohort to prevent longitudinal information leakage. The final composition is shown in Table 3.

Table 3: Composition of the longitudinal glioma MRI cohorts.

Dataset	Training	Validation
LUMIERE	397	92
MU	314	76
UCSF	238	60
Total	949	228

Image preprocessing. All longitudinal scans are registered to a common spatial reference and resampled to an isotropic spacing of $1.5\,\mathrm{mm}$. Volumes are then center cropped or padded to $120 \times 144 \times 120$ voxels. Image intensities are clipped using the 0.5th and 99.5th percentiles and normalized to $[0, 1]$. Lesion masks undergo the same spatial transformations.

Temporal and treatment conditions. The actual elapsed time between the paired examinations is retained as a continuous temporal condition. Treatment information is extracted from the corresponding clinical metadata and mapped to a common conditioning representation across cohorts.

Lesion definitions. Lesion annotations are converted to binary foreground masks after cohort-specific preprocessing. The underlying annotation semantics are not identical across datasets: LUMIERE foreground contains its annotated non-enhancing abnormality and enhancing tumor regions, whereas MU and UCSF additionally include the available resection-cavity label together with their tumor-related foreground classes. We therefore use a common binary foreground representation while retaining the native cohort-specific lesion definitions rather than imposing an artificial subregion correspondence.

B Training and Inference Implementation Details

Conditional diffusion backbone. Following BrLP Puglisi et al. (2025), the future image is encoded as $z_f = \mathcal{E}(X_f)$ and perturbed by standard forward diffusion,

$$z_t = \sqrt{\bar{\alpha}_t} z_f + \sqrt{1 - \bar{\alpha}_t}\epsilon. \tag{8}$$

The diffusion model predicts the injected noise under

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_f, t, \epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, \mathcal{C})\|_2^2\right], \tag{9}$$

where $\mathcal{C}$ contains the current MRI, elapsed time, and treatment. The image autoencoder and diffusion UNet are initialized from the pretrained forecasting model and remain frozen during tumor-guided fine-tuning.

ERM spatial inputs and initialization. From the current tumor mask M_s, we construct a signed distance map d_s clipped to $\pm 15\,\mathrm{mm}$ and a soft boundary map $b_s = \exp(-|d_s|/5\,\mathrm{mm})$. We also include the lesion-masked current MRI latent $M_s \odot z_s$, where $z_s = \mathcal{E}(X_s)$, yielding

$$G_s = [M_s, d_s, b_s, M_s \odot z_s]. \tag{10}$$

A lightweight spatial encoder maps G_s to multi-scale features aligned with the selected ControlNet residuals. The diffusion timestep, elapsed time, and treatment are jointly encoded as $e_{\text{evo}} = E_{\text{evo}}(t, \Delta t, c)$ and combined with the corresponding spatial feature to predict the affine parameters in Eq. 1. ERM is applied only to a sparse subset of residual scales, and its affine prediction layers are initialized so that the original ControlNet pathway is preserved at the start of fine-tuning.

Tumor segmenter and representation readout. We use a single fixed nnU-Net-based tumor segmenter implemented following the MoME framework Isensee et al. (2021); Zhang et al. (2025b). The segmenter is trained separately using T1C images from the training cohort and remains frozen throughout diffusion-model fine-tuning. The same checkpoint is used for ECL supervision and for the common image-based lesion evaluation applied to all forecasting methods. No method-specific auxiliary segmentation output is used when computing the reported Dice scores.

The representation readout used by ERA is pretrained before tumor-guided fine-tuning and is subsequently frozen. It therefore acts as a fixed readout of future-tumor information from intermediate generative features rather than as a jointly optimized auxiliary segmentation head.

Trajectory-scorer development. The prospective and terminal trajectory scorers are trained only from trajectories generated for patients in the training split. Scorer development uses patient-grouped five-fold out-of-fold predictions, ensuring that trajectories from the same patient do not appear in both scorer fitting and held-out folds. Once the scorer design and hyperparameters are fixed, the final scorers are refit using the complete training cohort.

The prospective scorer uses deployment-available descriptors derived from the partially denoised image hypothesis, predicted lesion structure, longitudinal change relative to the current examination, candidate-population agreement, and temporal and treatment conditions. The terminal scorer additionally uses tumor-related summaries obtained from the frozen ERA readout. Future MRI and future lesion annotations are used only to construct training-time trajectory preferences and are never provided as scorer inputs at inference.

Trajectory ranking and prospective selection. Training targets are constructed by segmenting completed predictions and measuring their Dice agreement q_k with the observed future mask. For a within-patient pair (i, j) with target $y_{ij} = \mathbf{1}[q_i > q_j]$, the scorer is trained with

$$\mathcal{L}_{\text{rank}} = -y_{ij} \log \sigma(s_i - s_j) - (1 - y_{ij}) \log \sigma(s_j - s_i). \tag{11}$$

At pruning stage ℓ, a one-step clean estimate is decoded from each surviving latent and segmented to obtain the intermediate lesion hypothesis. The prospective descriptor $\phi^{\text{pros}}_{k,\ell}$ summarizes this hypothesis together with longitudinal change, candidate-population agreement, image evidence, and clinical conditions. The scorer V_ℓ assigns $s^{\text{pros}}_{k,\ell} = V_\ell(\phi^{\text{pros}}_{k,\ell})$, after which

$$\mathcal{K}^{(\ell)} = \text{TopK}\left(\{s^{\text{pros}}_{k,\ell} : k \in \mathcal{K}^{(\ell-1)}\}, K_\ell\right). \tag{12}$$

Population-relative features are recomputed over the surviving candidates before each subsequent pruning operation. The future mask is used only to construct scorer-training preferences and is never supplied to the deployed scorers.

Sampling budget. All diffusion-based methods are evaluated with 50 sampling steps. Our prospective refinement follows

$$24 \xrightarrow{t=400} 20 \xrightarrow{t=200} 8. \tag{13}$$

Under the adopted sampling schedule, the corresponding nominal denoiser budget is

$$24 \times 30 + 20 \times 10 + 8 \times 10 = 1000 = 20 \times 50, \tag{14}$$

which is equivalent to completing 20 full trajectories. Diffusion-based baselines are therefore evaluated using 20 independent trajectories followed by uniform averaging.

After the eight surviving trajectories complete denoising, terminal scorer outputs are converted into soft aggregation weights using temperature $\tau = 1.0$.

C Baseline Implementation Details

C.1 SADM

SADM Yoon et al. (2023) is natively a three-dimensional longitudinal diffusion model; we therefore do not convert a 2D architecture into 3D. Instead, we adapt the official implementation to the pairwise T1C forecasting setting used in our experiments.

The original model accepts a sequence of historical 3D volumes. In our pairwise setting, only the designated current T1C volume is provided as context and the subsequent T1C volume serves as the prediction target. Thus, the 3D ViViT condition encoder is retained but receives a single historical frame. The original 3D ContextUNet, epsilon-prediction objective, diffusion process, timestep sampling, and conditioning mechanism are otherwise preserved.

Our common volumes have size $120 \times 144 \times 120$. To satisfy the spatial requirements of the released implementation, the first and third dimensions are symmetrically padded to obtain $128 \times 144 \times 128$, without resizing the image. Generated volumes are cropped back to the common $120 \times 144 \times 120$ evaluation space. The shape-dependent bottleneck pooling and upsampling operations in the released ContextUNet are generalized from their original hard-coded input geometry to the resulting $32 \times 36 \times 32$ bottleneck resolution.

For the ViViT condition encoder, the number of context frames is set to one and 3D patches of size $16 \times 16 \times 16$ are used. The released three-dimensional architecture is otherwise retained. Formal evaluation uses 50 sampling steps and Mean20 aggregation under the common denoising budget.

C.2 TADIFF-3D

TaDiff Liu et al. (2025) is adapted from its original slice-based formulation to full volumetric prediction while preserving its core multi-session, temporal, treatment-conditioned, joint MRI–segmentation framework.

Volumetric architecture. All two-dimensional spatial operations are replaced by their 3D counterparts. Convolutions use 3D kernels, downsampling and upsampling operate along all three spatial dimensions, and attention is applied after flattening the 3D feature volume into tokens. To control memory usage, attention is used only at the $\times 8$ downsampling level. The volumetric network uses $\texttt{channel_mult} = (1, 2, 3, 4)$ with four attention heads.

Single-modality channel configuration. The original configuration assumes multi-channel images. For T1C-only forecasting, each of the four session slots contains one MRI channel. The network therefore receives five input channels and produces five output channels: four tumor-mask logits, one for each session slot, and one diffusion noise channel for the target MRI.

Pairwise-to-multi-session mapping. Our data contain one current and one future examination, whereas the TaDiff interface expects four session slots. Each training pair is therefore represented as

$$[X_s, X_s, X_s, X_f], \tag{15}$$

with timestamps

$$[0, 0, 0, \Delta t]. \tag{16}$$

The first three slots repeat the same current examination only to preserve the original conditioning interface and do not represent three independent historical scans. Their treatment conditions correspond to the current state, while the final slot uses the actual treatment condition associated with the forecast interval. Training always treats the final slot as the prediction target.

Three-dimensional loss and sampling. The original tumor-weighted MRI objective and auxiliary segmentation supervision are retained and extended to full 3D volumes. Spatial weighting uses a $11 \times 11 \times 11$ kernel, and the segmentation Dice loss operates on volumetric predictions. The final objective is the weighted diffusion MRI loss plus 0.01 times the segmentation loss.

The model is trained on the same 949 training transitions with volumetric inputs of size $120 \times 144 \times 120$, using Adam with learning rate 2.5×10^{-4}, BF16 mixed precision, batch size one, and 600 diffusion timesteps. Training is run for up to 400k optimization steps. The formal evaluation uses the checkpoint at 190k steps, selected from validation performance, with 50-step DDIM sampling and Mean20 aggregation.

During reverse diffusion, the first three repeated current-image slots remain fixed as conditions and only the final future slot is iteratively updated. Generated volumes are written directly in the common 3D evaluation space.