GBM-WAM: ANATOMY-GUIDED LATENT WORLD-ACTION MODEL FOR BRAIN GLIOBLASTOMA TREATMENT PLANNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Planning next-step treatment for glioblastoma (GBM) requires current magnetic resonance imaging (MRI) and clinical history, while follow-up scans provide longitudinal supervision for modeling treatment-relevant progression beyond direct observation-to-action prediction. World-action models (WAMs) provide a natural paradigm for exploiting such longitudinal information, yet common imagine-then-act designs generate explicit future observations before predicting the action. For volumetric MRI, explicit future generation requires costly reconstruction of dense 3D anatomy and appearance details that are not uniformly relevant to treatment, while treatment-relevant progression is distributed across complementary MRI sequences that may not all be available in clinical practice. We therefore propose GBM-WAM, an anatomy-guided latent WAM that recasts explicit future volume generation as partial-to-complete latent world learning within a causal multimodal large language model (MLLM), enabling training-rich, inference-light planning from variable MRI subsets and clinical context. To learn GBM progression dynamics from partial MRI inputs, a latent world expert combines a mixture-of-transformer architecture with flow matching to predict complete multi-sequence future representations, shaping world query states with shared and sequence-specific progression cues. Concurrently, a temporal mask expert performs state-modulated spatial decoding to supervise localized tumor morphology and longitudinal volumetric changes, providing complementary anatomical guidance. While both auxiliary experts supervise intermediate states during training, they are bypassed at inference to enable direct treatment planning in a single MLLM pass. Experiments on internal and external multicenter cohorts show that GBM-WAM outperforms state-of-the-art methods, achieving up to 93.2% accuracy and up to 91.6% F1 score, while accommodating variable combinations of available MRI sequences and requiring only 0.375 s per case for direct treatment inference.

1 INTRODUCTION

Planning next-step treatment for glioblastoma (GBM) requires integrating evidence beyond a single clinical visit (Wen et al., 2020; Weller et al., 2021). Current magnetic resonance imaging (MRI) and clinical history characterize the disease state and its prior trajectory, while longitudinal assessment provides complementary evidence about treatment response, recurrence, and disease evolution (Wang et al., 2016; Wen et al., 2023). Consequently, follow-up imaging offers valuable supervision for learning GBM progression dynamics that inform treatment prediction.

World-action models (WAMs) (Cen et al., 2025; Ye et al., 2026; Yuan et al., 2026) provide a natural paradigm for exploiting such longitudinal supervision by modeling predictive state transitions that inform action selection, beyond a direct *observation-to-action* mapping. A common realization is the *imagine-then-act* paradigm, in which the model first predicts an explicit future observation and subsequently derives the action from that predicted state, making explicit future prediction integral to both training and inference. For volumetric MRI, this prerequisite is particularly costly because observation-level forecasting must

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.

reconstruct an entire dense, high-dimensional 3D volume, including large regions of anatomy and fine-grained appearance details that contribute little to treatment planning. This limitation motivates investigating whether WAMs can learn from future progression without making explicit future MRI volume generation a prerequisite for treatment prediction.

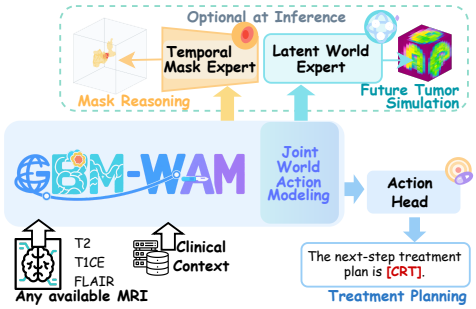


Figure 1: **Overview of GBM-WAM.** GBM-WAM learns future latent dynamics and temporal anatomical cues via auxiliary training supervision for direct treatment planning.

Predicting future states in compact latent spaces offers a scalable alternative to high-dimensional observation synthesis (Zhou et al., 2025; Baldassarre et al., 2025), but naively compressing 3D scans into compact latents risks discarding treatment-dependent responses and subtle subregional tumor changes essential for clinical decisions. In GBM, treatment-relevant progression is distributed across complementary MRI sequences, including T2-weighted (T2), contrast-enhanced T1-weighted (T1CE), and fluid-attenuated inversion recovery (FLAIR), each revealing distinct aspects of tumor pathology. However, heterogeneous acquisition protocols often leave one or more sequences unavailable, such that the observed inputs provide only a partial view of disease evolution. Most imaging-based GBM prognostic models remain focused on single-visit assessments (Suter et al., 2022; Luckett et al., 2023; Akbari et al., 2025), whereas incomplete-modality methods (Li et al., 2025; Zeng et al., 2024) address missing sequences without modeling treatment-conditioned evolution across arbitrary MRI subsets. Consequently, the central challenge is to avoid explicit observation synthesis by learning compact predictive latents that recover sequence-complete future representations from partial MRI inputs while preserving localized, treatment-dependent tumor dynamics and directly supporting treatment planning.

To address this challenge, we propose **GBM-WAM**, an anatomy-guided latent world-action model that integrates GBM world modeling and treatment planning within a single causal multimodal large language model (MLLM), enabling training-rich, inference-light prediction from variable MRI subsets together with clinical context. At its core, as shown in Figure 1, GBM-WAM formulates world learning as partial-to-complete latent prediction, shaping progression-aware world query states that directly support action prediction. Specifically, a latent world expert predicts complete multi-sequence future representations from these states, while a temporal mask expert provides complementary supervision to the preceding mask query states by capturing localized tumor morphology at the current and future time points, together with the longitudinal change between them. Current available MRI and clinical context determine what the model observes, whereas the longitudinal latent and anatomical targets enrich the intermediate states with complementary progression and structural cues. At inference, GBM-WAM retains the supervised mask and world query states but bypasses both auxiliary experts, predicting treatment directly in a single causal MLLM pass.

The main contributions of this work are summarized as follows.

1. We propose GBM-WAM, an anatomy-guided latent world-action model that performs treatment planning from variable MRI subsets within a causal MLLM, uses a future-state conditioner to restrict future conditioning to auxiliary training, and directly predicts treatment without auxiliary decoding at inference.

2. We design a latent world expert that combines a mixture-of-transformers architecture with flow matching to predict complete multi-sequence future representations from partial MRI inputs, shaping world states with shared and sequence-specific tumor progression cues.

3. We design a temporal mask expert with state-modulated spatial decoding that supervises mask query states through localized tumor morphology at current and future time points and longitudinal changes between them, providing complementary anatomical guidance.

4. Experiments on internal and external multicenter GBM cohorts show that GBM-WAM outperforms state-of-the-art methods, achieving accuracy/F1 scores of 93.2%/91.6% internally and 81.9%/78.2% externally, while supporting all non-empty MRI subsets and requiring only 0.375 s per case for direct treatment inference.

2 Related Work

GBM segmentation, grading, and treatment modeling. GBM evolves rapidly, making longitudinal MRI essential for diagnosis and treatment planning, yet most models remain limited to single-visit analysis. Multi-parametric benchmarks standardize sub-region delineation (Bonato et al., 2025), while transformer-based and foundation architectures advance 3D segmentation (Hatamizadeh et al., 2021; Ma et al., 2024). Other studies combine segmentation with GBM grading (Naser & Deen, 2020; Zhuge et al., 2020). Despite their clinical value, these methods neither model treatment-conditioned evolution nor guide subsequent interventions. Longitudinal approaches therefore extend tumor characterization toward treatment planning. CLARITY models treatment-conditioned disease trajectories for decision optimization (Ding et al., 2026), while Brain-WM jointly predicts the next-step treatment and generates future MRI within a shared latent space (Wang et al., 2026). Our GBM-WAM differs in its partial-to-complete learning setup: it learns anatomy-guided latent worlds from arbitrary current MRI subsets, using modality-complete future targets to shape causally ordered mask and world states.

Predictive representations and world-action models. World models (Ha & Schmidhuber, 2018; LeCun, 2022) learn compact states and action-conditioned transitions for planning. JEPA (Assran et al., 2023) predicts task-relevant embeddings rather than reconstructing future observations. DINO-WM (Zhou et al., 2025) predicts pretrained features for planning, while JEPA-style NeuroVFM (Kondepudi et al., 2026) provides anatomy-aware representations learned from 5.24 million CT and MRI volumes. Vision-language-action models connect multimodal perception with action (Zitkovich et al., 2023; Kim et al., 2025), and WAMs show that learning future dynamics can improve action decisions but often require explicit future prediction at inference. This raises a question: *Must future prediction run at inference?* WorldVLA (Cen et al., 2025) unifies image and action tokens within an autoregressive model, whereas Fast-WAM (Yuan et al., 2026) couples a video DiT observation encoder with an action-denoising expert and skips explicit future prediction at inference. In GBM, Brain-WM (Wang et al., 2026) shares representations across planning and imaging branches, whereas GBM-WAM predicts treatment from temporally supervised mask states and future-supervised world states via a causal MLLM, without auxiliary decoding at inference.

3 Methodology

3.1 Problem Formulation and Overview

As illustrated in Figure 2, GBM-WAM is an anatomy-guided latent WAM that integrates world modeling and treatment planning within a single causal MLLM, with temporal mask reasoning providing complementary anatomical guidance for training-rich, inference-light treatment planning.

Problem formulation. At clinical visit t, GBM-WAM observes the available MRI volumes $\boldsymbol{x}_t = \{\boldsymbol{x}_t^m \mid m \in \mathcal{M}, v_t^m = 1\}$, where $\mathcal{M} = \{\mathrm{T2}, \mathrm{T1CE}, \mathrm{FLAIR}\}$ and $\boldsymbol{v}_t \in \{0,1\}^3 \setminus \{\mathbf{0}\}$ indicates modality availability ($v_t^m = 1$ if modality m is observed), together with clinical context $\boldsymbol{c}_{\le t}$, comprising patient demographics and prior treatment history. Treatments are represented by five newly added special tokens, $\mathcal{A} = \{\texttt{CRT}, \texttt{RT}, \texttt{TMZ}, \texttt{SUR}, \texttt{AM}\}$, denoting chemoradiotherapy, radiotherapy, temozolomide chemotherapy, surgery, and active monitoring (no intervention), respectively. Our task is to predict the next-step treatment intervention along the longitudinal trajectory conditioned on multimodal MRI and clinical context, serving as a *methodological proof of concept* to evaluate GBM-WAM for adaptive treatment planning. Given the model's conditional distribution p over these tokens, we formulate the treatment-planning task as:

$$a_{t+1}^* = \arg\max_{a \in \mathcal{A}} \; p(a \mid \boldsymbol{x}_t, \boldsymbol{c}_{\le t}). \tag{1}$$

Flexible multi-sequence input. To handle missing MRI modalities, we represent each missing sequence with a newly added special token <MRI_seq>. For each available sequence, we independently pass its MRI volume through the same frozen NeuroVFM encoder $E(\cdot)$ (Kondepudi et al., 2026), pretrained on large-scale clinical neuroimaging data, to extract anatomy-aware visual tokens. We arrange the modality slots in the fixed order T2, T1CE, and FLAIR, insert <MRI_seq> at each missing slot, and place the resulting MRI token sequence before the clinical context tokens in the MLLM's input, providing a consistent structure for any non-empty MRI subset.

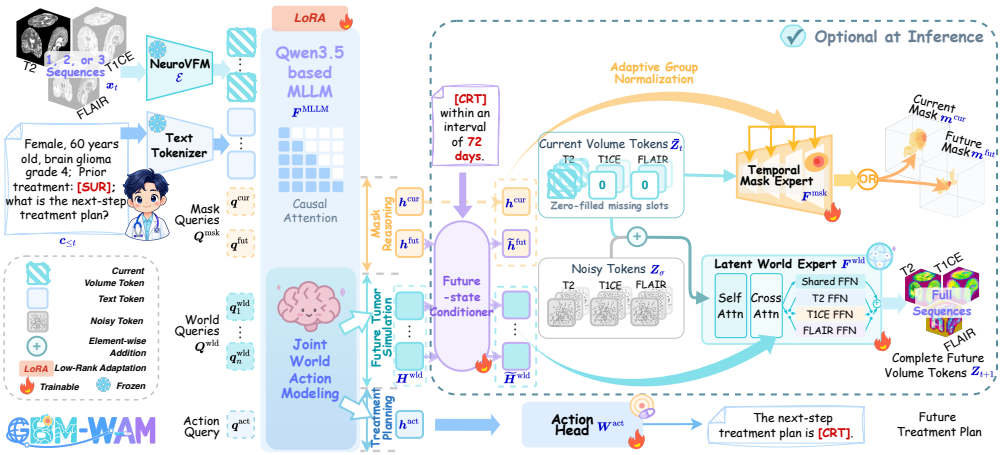


Figure 2: **Illustration of GBM-WAM.** The causal MLLM ingests arbitrary MRI subsets with patient context via ordered queries. Latent world and temporal mask experts regularize intermediate states during training, while inference predicts treatment in one forward pass without auxiliary decoding.

Joint world-action modeling. To connect tumor dynamics with treatment planning, we construct a query sequence comprising N_{wld} world queries $\boldsymbol{Q}^{\text{wld}} = [\boldsymbol{q}_i^{\text{wld}}]_{i=1}^{N_{\text{wld}}}$ followed by an action query $\boldsymbol{q}^{\text{act}}$. A single causal MLLM pass produces their corresponding hidden states $\boldsymbol{H}^{\text{wld}}$ and $\boldsymbol{h}^{\text{act}}$, allowing the action query to attend to preceding world states. The first two expressions below describe this within-pass dependency, followed by treatment prediction through an action head $\boldsymbol{W}^{\text{act}}$:

$$\begin{aligned} \boldsymbol{H}^{\text{wld}} &= F^{\text{MLLM}}\left[\boldsymbol{Q}^{\text{wld}} \mid E(\boldsymbol{x}_t), \boldsymbol{c}_{\le t}\right], \\ \boldsymbol{h}^{\text{act}} &= F^{\text{MLLM}}\left[\boldsymbol{q}^{\text{act}} \mid E(\boldsymbol{x}_t), \boldsymbol{c}_{\le t}, \boldsymbol{H}^{\text{wld}}\right], \\ p(a \mid \boldsymbol{x}_t, \boldsymbol{c}_{\le t}) &= \frac{\exp\left([\boldsymbol{W}^{\text{act}}\boldsymbol{h}^{\text{act}}]_a\right)}{\sum_{a' \in \mathcal{A}} \exp\left([\boldsymbol{W}^{\text{act}}\boldsymbol{h}^{\text{act}}]_{a'}\right)}. \end{aligned} \tag{2}$$

To support current and future mask reasoning and provide complementary anatomical cues, we prepend mask queries $\boldsymbol{Q}^{\text{msk}} = [\boldsymbol{q}^{\text{cur}}, \boldsymbol{q}^{\text{fut}}]$ to the world-action input sequence and obtain hidden states $\boldsymbol{h}^{\text{cur}}$ and $\boldsymbol{h}^{\text{fut}}$ in the same MLLM forward pass. During training, the future-state conditioner incorporates the factual treatment and follow-up interval into the post-pass world hidden states $\boldsymbol{H}^{\text{wld}}$ and future-mask state $\boldsymbol{h}^{\text{fut}}$, yielding $\widetilde{\boldsymbol{H}}^{\text{wld}}$ and $\widetilde{\boldsymbol{h}}^{\text{fut}}$. Conditioned on $\widetilde{\boldsymbol{H}}^{\text{wld}}$, a latent world expert F^{wld} predicts the modality-complete future MRI token sequence $\boldsymbol{Z}_{t+1} = [E(\boldsymbol{x}_{t+1}^m)]_{m \in \mathcal{M}}$ via flow matching for arbitrary non-empty current MRI subsets, encouraging world states to encode cross-sequence tumor dynamics relevant to treatment planning. In parallel, a temporal mask expert F^{msk} decodes $\boldsymbol{h}^{\text{cur}}$ and $\widetilde{\boldsymbol{h}}^{\text{fut}}$ into the current and future GBM masks, $\boldsymbol{m}_t$ and $\boldsymbol{m}_{t+1}$, respectively, encouraging the mask states to represent tumor extent and temporal change.

Inference-light treatment planning. At inference, GBM-WAM bypasses auxiliary experts F^{wld} and F^{msk} while retaining the mask and world query states. Training-time world modeling and anatomical reasoning shape these states to encode implicit cues about tumor progression and structure. Thus, GBM-WAM directly predicts the next-step treatment a_{t+1} in a single causal MLLM pass, retaining the benefit of auxiliary supervision without running the experts at inference.

3.2 Future-State Treatment-Time Conditioning

Treatment planning uses only evidence available at decision time, whereas auxiliary future-state prediction is conditioned on the treatment course $r_{t:t+1}$ and follow-up interval Δt. We define $r_{t:t+1}$ as the interventions recorded between baseline visit t and its paired follow-up visit $t+1$. During training, we encode the factual pair $(r_{t:t+1}, \Delta t)$ as a structured prompt using the shared MLLM embedding table, yielding $\boldsymbol{C}$. For each future-facing representation $\boldsymbol{U} \in \{\boldsymbol{H}^{\text{wld}}, \boldsymbol{h}^{\text{fut}}\}$, we apply a future-state conditioner based on gated cross-attention:

$$\widetilde{\boldsymbol{U}} = \boldsymbol{U} + \tanh(\alpha)\,\text{MHCA}\left(\text{LN}(\boldsymbol{U}), \text{LN}(\boldsymbol{C}), \text{LN}(\boldsymbol{C})\right), \tag{3}$$

where LN is layer normalization, MHCA is multi-head cross-attention, and α is a zero-initialized learnable scalar. Future-state conditioning is applied only after the causal MLLM pass for auxiliary future MRI and mask prediction, while the action query accesses only the unconditioned states.

3.3 MoT-Based Latent World Prediction

To provide prognostic supervision, the latent world expert F^{wld} predicts future full-sequence MRI tokens conditioned on $\widetilde{\boldsymbol{H}}^{\mathrm{wld}}$. Instead of reconstructing voxel intensities, we define the target in latent space as $\boldsymbol{Z}_{t+1} = [E(\boldsymbol{x}_{t+1}^m)]_{m\in\mathcal{M}}$. Requiring this complete target regardless of input availability encourages the world states to encode future tumor dynamics and cross-sequence relationships.

Mixture-of-transformers (MoT) architecture. To capture shared structure and sequence-specific features, we implement F^{wld} as a linear-attention transformer (Xie et al., 2025a) whose 3D feed-forward blocks combine shared and sequence-specific branches. For sequence $m \in \mathcal{M}$, let $\boldsymbol{Y}^m$ denote its block tokens. The MoT feed-forward network (FFN) computes:

$$\mathrm{FFN}_{\mathrm{MoT}}(\boldsymbol{Y}^m) = \mathrm{FFN}_{\mathrm{shared}}(\boldsymbol{Y}^m) + \mathrm{FFN}_m(\boldsymbol{Y}^m), \tag{4}$$

where $\mathrm{FFN}_{\mathrm{shared}}$ uses shared parameters, while FFN_m uses sequence-specific parameters.

Partial-to-complete latent world prediction. We formulate future latent prediction through rectified flow matching. Specifically, we sample Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and a flow time $\sigma \sim \mathcal{U}(0,1)$, defining the interpolated state $\boldsymbol{Z}_\sigma$ and target velocity $\boldsymbol{u}_\sigma$ as:

$$\boldsymbol{Z}_\sigma = (1-\sigma)\boldsymbol{Z}_{t+1} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{u}_\sigma = \boldsymbol{\epsilon} - \boldsymbol{Z}_{t+1}. \tag{5}$$

Unlike the variable-length MLLM input with special tokens for missing modalities, the auxiliary experts' spatial inputs retain fixed sequence slots and grids aligned with the complete future target. We therefore construct $\bar{\boldsymbol{Z}}_t = [\bar{\boldsymbol{Z}}_t^{\mathrm{T2}}, \bar{\boldsymbol{Z}}_t^{\mathrm{T1CE}}, \bar{\boldsymbol{Z}}_t^{\mathrm{FLAIR}}]$, filling each slot with current NeuroVFM tokens if $v_t^m = 1$ and zeros otherwise. Element-wise addition of $\bar{\boldsymbol{Z}}_t$ to $\boldsymbol{Z}_\sigma$ provides spatial conditioning, while cross-attention to $\widetilde{\boldsymbol{H}}^{\mathrm{wld}}$ supplies treatment-conditioned prognostic context. Given both conditions, the world expert estimates the velocity field and minimizes the flow-matching loss:

$$\widehat{\boldsymbol{u}}_\sigma = F^{\mathrm{wld}}\big(\boldsymbol{Z}_\sigma + \bar{\boldsymbol{Z}}_t, \sigma, \widetilde{\boldsymbol{H}}^{\mathrm{wld}}\big), \qquad \mathcal{L}_{\mathrm{world}} = \mathbb{E}_{\sigma,\boldsymbol{\epsilon}}[\big\|\widehat{\boldsymbol{u}}_\sigma - \boldsymbol{u}_\sigma\big\|_2^2]. \tag{6}$$

Flow matching diversifies training across noise realizations and interpolation times, while MoT combines shared and sequence-specific transformations for multi-sequence future prediction.

3.4 State-Modulated Temporal Mask Reasoning

To provide anatomical cues before treatment planning, the temporal mask expert F^{msk} first performs current and future mask reasoning by decoding one mask at a time from current volume tokens $\bar{\boldsymbol{Z}}_t$, guided by $\boldsymbol{h}^{\mathrm{cur}}$ for current GBM mask prediction or $\widetilde{\boldsymbol{h}}^{\mathrm{fut}}$ for future GBM mask prediction. A multi-scale hierarchical backbone extracts these features via 3D linear self-attention, with binary gating suppressing missing-sequence features according to $\boldsymbol{v}_t$. At each fusion stage, sequence-specific features are concatenated with upsampled features from the preceding scale and fused via a convolutional projection. The selected temporal state progressively modulates these features to guide the corresponding segmentation.

Temporal state modulation. To support both timepoints with a shared decoder, we use state-conditioned adaptive group normalization. A linear projection maps the selected input state to a temporal control vector $\boldsymbol{g}^k$, where $k \in \{\mathrm{cur}, \mathrm{fut}\}$ identifies the prediction timepoint. Learned mappings $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ transform this vector into scale and shift parameters that modulate normalized decoder features. Given intermediate features $\boldsymbol{B}^k$, each residual refinement block applies

$$\widetilde{\boldsymbol{B}}^k = \boldsymbol{B}^k + \boldsymbol{\gamma}(\boldsymbol{g}^k) \odot \mathrm{GN}(\boldsymbol{B}^k) + \boldsymbol{\beta}(\boldsymbol{g}^k), \tag{7}$$

where GN is group normalization and $\odot$ is element-wise multiplication. This transformation adapts decoder features to the selected timepoint, producing mask logits $\boldsymbol{G}^k$ for background and GBM subregions: edema (ED), enhancing tumor (ET), and non-enhancing tumor core (NET).

Longitudinal change supervision. We combine per-visit segmentation supervision with a voxel-wise change objective to capture temporal transitions. The change score $\widehat{\boldsymbol{d}}(p)$ measures the difference between current and future predicted class distributions and is supervised against the binary target $\boldsymbol{d}(p)$, which indicates whether the ground-truth class changes at voxel p. Using focal loss $\mathcal{L}_{\text{focal}}$, Dice loss $\mathcal{L}_{\text{Dice}}$, and binary cross-entropy loss $\mathcal{L}_{\text{BCE}}$, we define the joint objective:

$$\begin{aligned} \mathcal{L}_{\text{seg}}(\boldsymbol{G}, \boldsymbol{m}) &= \mathcal{L}_{\text{focal}}(\boldsymbol{G}, \boldsymbol{m}) + \mathcal{L}_{\text{Dice}}(\boldsymbol{G}, \boldsymbol{m}), \\ \widehat{\boldsymbol{d}}(p) &= \tfrac{1}{2} \left\| \text{softmax}(\boldsymbol{G}^{\text{fut}}(p)) - \text{softmax}(\boldsymbol{G}^{\text{cur}}(p)) \right\|_1, \\ \mathcal{L}_{\text{mask}} &= \underbrace{\mathcal{L}_{\text{seg}}(\boldsymbol{G}^{\text{cur}}, \boldsymbol{m}_t)}_{\text{Current segmentation}} + \underbrace{\mathcal{L}_{\text{seg}}(\boldsymbol{G}^{\text{fut}}, \boldsymbol{m}_{t+1})}_{\text{Future segmentation}} + \underbrace{\mathcal{L}_{\text{BCE}}(\widehat{\boldsymbol{d}}, \boldsymbol{d}) + \mathcal{L}_{\text{Dice}}(\widehat{\boldsymbol{d}}, \boldsymbol{d})}_{\text{Voxel-wise change}}. \end{aligned} \tag{8}$$

3.5 Joint Training and Inference-Light Prediction

To incorporate anatomical and prognostic cues into treatment planning, the action query follows and causally attends to the mask states $[\boldsymbol{h}^{\text{cur}}, \boldsymbol{h}^{\text{fut}}]$ and world states $\boldsymbol{H}^{\text{wld}}$, which remain unconditioned on factual future treatment and interval information along the treatment planning path. During training, we apply sequence dropout to current inputs from complete three-sequence MRI data, retaining at least one sequence and keeping future targets complete.

We reuse the MLLM language-modeling head as $\boldsymbol{W}^{\text{act}}$ and normalize the corresponding logits over the treatment-token set $\mathcal{A}$. The target treatment token a_{t+1} is supervised through cross-entropy:

$$\mathcal{L}_{\text{act}} = -\log \frac{\exp\left([\boldsymbol{W}^{\text{act}} \boldsymbol{h}^{\text{act}}]_{a_{t+1}}\right)}{\sum_{a' \in \mathcal{A}} \exp\left([\boldsymbol{W}^{\text{act}} \boldsymbol{h}^{\text{act}}]_{a'}\right)}. \tag{9}$$

While this objective directly supervises treatment planning, the world and mask objectives shape the preceding query states through future-representation prediction and temporal mask reasoning. With auxiliary loss weights λ_{world} and λ_{mask}, the joint objective is:

$$\mathcal{L} = \mathcal{L}_{\text{act}} + \lambda_{\text{world}} \mathcal{L}_{\text{world}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}}. \tag{10}$$

At inference, a single causal MLLM pass maps observed MRI and clinical history to mask, world, and action states, from which GBM-WAM selects the highest-probability treatment token in $\mathcal{A}$ without flow integration or mask decoding. For optional visualization, auxiliary experts can generate future MRI latents and masks conditioned on a specified treatment and follow-up interval.

4 Experiments

4.1 Experimental Setup

Datasets and preprocessing. The internal cohort combines LUMIERE (Suter et al., 2022), MU-Glioma Post (Mahmoud et al., 2025), and UCSF-ALPTDG (Fields et al., 2024), comprising 527 subjects, 1,659 timepoints, and 3,252 longitudinal pairs, with an 80:20 patient-level split to prevent leakage. UCSD-PTGBM (Gagnon et al., 2026) provides 49 subjects and 83 pairs for external zero-shot evaluation, with its predominance of repeat-surgery cases posing an additional generalization challenge. Across cohorts, we retained subjects with at least two available timepoints and derived pairwise longitudinal samples. Each sample includes available baseline T2, T1CE, and FLAIR volumes and clinical context for treatment prediction. Scans are skull-stripped, co-registered to SRI24 template space (Rohlfing et al., 2010), and cropped to $140 \times 192 \times 192$. Cohort and treatment distributions are detailed in Appendix A, while the five-way task rationale is discussed in Appendix B.

Implementation details and evaluation metrics. We connect frozen NeuroVFM (Kondepudi et al., 2026) to the causal MLLM Qwen3.5-0.8B (Qwen Team, 2026) via a visual adapter. NeuroVFM's pretraining corpus, UM-NeuroImages, comprises clinical imaging from michigan medicine and has no dataset-level overlap with our experimental cohorts (Kondepudi et al., 2026). Two temporal mask queries, 128 world queries, and an action query follow the MRI and clinical tokens. We apply rank-96 LoRA (Hu et al., 2022) to the MLLM and jointly train new token embeddings and tied output-head rows, query embeddings, the future-state conditioner, and both auxiliary

Table 1: **Quantitative comparison on internal and external cohorts.** Baselines: open-source VLMs (I), closed-source VLMs (II), medical VLMs (III), traditional ML (IV), task-specific networks (V), and future simulation models (VI). Best results are highlighted in **bold**.

Method	Type	*Internal Cohort*				*External Cohort*			
		Accuracy ↑	F1-Score ↑	Specificity ↑	Precision ↑	Accuracy ↑	F1-Score ↑	Specificity ↑	Precision ↑
Qwen3-VL 4B	I	20.1	27.7	80.2	52.9	34.9	29.6	84.0	46.9
InternVL3.5 4B	I	30.2	28.4	80.0	31.3	32.5	24.2	83.8	29.4
MinimaxM3	I	35.3	29.8	84.7	63.7	32.5	21.4	84.0	34.8
Claude Haiku 4.5	II	29.3	40.2	85.5	91.1	28.9	16.0	84.1	11.3
Gemini 2.5 Pro	II	32.2	38.8	84.3	67.8	35.0	30.4	83.5	42.6
GPT-5.1	II	36.1	45.1	85.8	82.1	41.0	36.9	85.1	48.0
MedVLM-R1 2B	III	20.5	25.9	80.7	46.9	15.7	9.7	78.7	21.8
Hulu-Med 4B	III	24.6	27.5	82.2	50.9	14.5	13.8	79.4	23.0
Lingshu 7B	III	32.7	43.0	85.5	79.4	26.5	17.3	80.4	23.2
Logistic Regression	IV	74.2	62.7	92.1	58.7	54.2	47.0	87.6	47.5
Random Forest	IV	77.1	68.0	92.7	64.6	61.4	52.4	89.5	53.0
XGBoost	IV	77.8	69.0	92.9	65.6	60.2	51.1	89.2	51.4
AliFuse	V	81.6	79.8	89.8	86.3	62.7	56.4	89.9	56.3
HyperFusion	V	85.6	80.0	91.8	82.1	69.9	62.4	91.8	63.3
Show-O2^{ft}	VI	87.2	88.6	92.4	88.4	73.5	68.5	92.8	70.4
Brain-WM	VI	91.5	90.4	95.8	90.7	78.3	71.6	94.2	73.7
GBM-WAM (ours)		**93.2**	**91.6**	**97.5**	**92.0**	**81.9**	**78.2**	**95.1**	**79.8**

Note: ^{ft} denotes that Show-O2 is fine-tuned on the same training data and task using its original model architecture.

experts. We train for 80,000 AdamW steps with a 10^{-4} initial learning rate, 200-step warm-up, and cosine decay. Both auxiliary loss weights equal one ($\lambda_{\text{world}} = 1$ and $\lambda_{\text{mask}} = 1$). Treatment planning retains mask and world query states while bypassing auxiliary experts and flow integration; optional latent visualization uses 50-step Heun sampling (Karras et al., 2022). We evaluate treatment predictions against recorded treatments using accuracy, F1-score, specificity, and precision.

4.2 Comparison with SOTA methods

Using complete three-sequence MRI and clinical context, we compare GBM-WAM with six baseline categories on both the internal and external datasets, including vision-language models (VLMs): open-source general VLMs (I: Qwen3-VL 4B (Bai et al., 2025), InternVL3.5 4B (Wang et al., 2025), MinimaxM3 (Lai et al., 2026)), closed-source frontier VLMs (II: Claude Haiku 4.5 (Anthropic, 2025), Gemini 2.5 Pro (Comanici et al., 2025), GPT-5.1 (OpenAI, 2025)), and medical VLMs (III: MedVLM-R1 2B (Pan et al., 2025), Hulu-Med 4B (Jiang et al., 2025), Lingshu 7B (LASA Team et al., 2025)). The remaining categories are traditional machine learning methods (IV: Logistic Regression, Random Forest, XGBoost), task-specific deep networks (V: AliFuse (Chen & Hong, 2024), HyperFusion (Duenias et al., 2025)), and recent architectures incorporating future simulation (VI: fine-tuned Show-O2, denoted Show-O2^{ft} (Xie et al., 2025b), and Brain-WM (Wang et al., 2026)). We evaluate I–III off-the-shelf and train IV–VI on the same data for task-specific evaluation. In particular, Show-O2^{ft} serves as a fine-tuned unified multimodal model to isolate domain-adaptation gains in a large model. For VLMs (I–III), we serialize uniformly sampled axial MRI slices as video-like sequences in a fixed sequence order, with identical maximum slice-sampling settings. Traditional machine learning baselines (IV) use hand-engineered tabular features comprising clinical demographics, physical tumor-subregion volumes, and subregion tumor-burden fractions. Task-specific deep networks and future simulation models (V–VI) use 3D encoders.

The results in Table 1 highlights the challenges of translating multimodal understanding into GBM treatment planning. *First*, although closed-source VLMs (II) slightly outperform open-source VLMs (I), both remain below 42% accuracy. Medical VLMs (III) offer no consistent improvement, with external accuracy falling as low as 14.5%. Together, these results highlight that off-the-shelf VLMs lack the domain reasoning needed for treatment planning. *Second*, traditional machine learning models (IV) achieve up to 77.8% internal accuracy using explicit tumor-subregion volumes and burden fractions, highlighting the value of task-relevant anatomical information. This advantage, however, comes with reliance on test-time tumor annotations and scalar summaries that discard detailed 3D spatial structure. *Third*, task-specific deep networks (V) address these limitations by learning volumetric features directly from MRI and integrating clinical context, improving treatment planning without hand-engineered tumor summaries. Their snapshot-to-action formulation nevertheless

Table 2: **GBM-WAM performance across all seven non-empty MRI subsets on the internal cohort.** Δ: percentage-point (pp) gain over the T2-only baseline.

Modalities			Modality	Accuracy ↑		F1-Score ↑		Specificity ↑		Precision ↑	
T2	T1CE	FLAIR	Number	%	Δ	%	Δ	%	Δ	%	Δ
✓			1	87.4	–	84.6	–	94.8	–	85.1	–
	✓			88.6	+1.2	86.2	+1.6	95.3	+0.5	86.9	+1.8
		✓		89.9	+2.5	87.8	+3.2	95.9	+1.1	88.2	+3.1
✓	✓		2	89.6	+2.2	87.5	+2.9	95.8	+1.0	88.0	+2.9
✓		✓		91.5	+4.1	89.6	+5.0	96.6	+1.8	89.9	+4.8
	✓	✓		92.3	+4.9	90.7	+6.1	97.0	+2.2	91.1	+6.0
✓	✓	✓	3	**93.2**	**+5.8**	**91.6**	**+7.0**	**97.5**	**+2.7**	**92.0**	**+6.9**

leaves longitudinal progression unmodeled, motivating the incorporation of future dynamics. *Finally*, while task-specific fine-tuning boosts large models like Show-O2$^{\text{ft}}$ (VI) to 87.2% internal accuracy, GBM-WAM combines latent world-action modeling with temporal anatomical guidance to outperform all evaluated baselines across all metrics, achieving 93.2% internal and 81.9% external accuracy. These findings confirm that modeling future dynamics provides essential inductive biases beyond standard fine-tuning, with benefits extending to the external cohort.

4.3 Flexible Treatment Planning Across Arbitrary MRI Subsets

On the internal cohort, we evaluate whether GBM-WAM maintains treatment planning with limited MRI inputs and benefits from additional sequences across all seven non-empty sequence subsets (Table 2). *First*, single-sequence inputs achieve 87.4%–89.9% accuracy without architectural changes. Although T1CE is central to assessing enhancing disease in GBM (Wen et al., 2023), FLAIR performs best individually in this task, suggesting that treatment-relevant information extends beyond enhancement. *Second*, T1CE+FLAIR achieves the best pairwise performance (92.3% accuracy and 90.7% F1-score), consistent with the complementary depiction of enhancing lesions and surrounding tissue abnormalities, including edema and nonenhancing tissue changes (Ellingson et al., 2015). *Finally*, adding T2 further improves accuracy and F1-score to 93.2% and 91.6%, exceeding the T2-only baseline by 5.8 and 7.0 percentage points, respectively. These results demonstrate flexible planning across sequence subsets and effective use of complementary sequences when available.

4.4 Effect of Auxiliary Experts

Table 3 presents a progressive component ablation on the internal cohort, first evaluating future latent supervision without mask reasoning (Idxs. 1–4), and then assessing mask reasoning on a fixed world-supervised configuration (Idxs. 4–7). From the action-only baseline, direct MSE regression of future latents improves the F1-score from 83.2% to 85.9%. Replacing it with flow-matching supervision further improves accuracy from 88.6% to 90.1% and F1-score from 85.9% to 87.8%, supporting flow matching as an auxiliary treatment-prediction objective. Adding sequence-specific MoT branches raises the F1-score to 89.2% (Idx. 4), showing the benefit of sequence-specific modeling in the latent world expert. On this configuration, current and future mask reasoning improve the F1-score to 90.4% and 90.7%, respectively (Idxs. 5–6), showing that both anatomical targets help. The complete temporal mask objective achieves the highest F1-score of 91.6% (Idx. 7), supporting joint supervision of current anatomy and longitudinal structure. These results support latent world-action learning with anatomical guidance, while both auxiliary experts bypassed at inference.

To illustrate spatial information learned through auxiliary objectives, Figure 3 presents a case where GBM-WAM correctly predicts surgery (`[SUR]`) as the first intervention after the current visit. For auxiliary visualization, we condition future latent prediction and mask generation on the factual treatment course `[SUR]+[CRT]` and a 252-day follow-up interval. Here, `[SUR]` is the next-step action target, while `[SUR]+[CRT]` records the treatment course, including chemoradiotherapy before the follow-up scan. Following PCA-based latent visualization in DINO (Oquab et al., 2024; Siméoni et al., 2026), we fit a shared PCA basis across timepoints for each sequence and map the leading three components to RGB with common normalization. In the predicted FLAIR latents, the tumor-associated pattern largely resolves, whereas the affected region in the T2W latents shrinks, consistent with changes following surgery and chemoradiotherapy. Together with the predicted masks, these patterns illustrate how the auxiliary objectives capture both progression-related latents and explicit tumor structure, complementing the quantitative ablation results.

Table 3: **Component ablation of GBM-WAM on the internal cohort.** Δ: percentage-point (pp) gain over the core baseline (Idx. 1). Best results are **bold**.

Idx.	*Action Head*	*Latent World Expert*		*Temporal Mask Expert*		Accuracy ↑		F1-Score ↑		Specificity ↑		Precision ↑	
		Co-training	MoT	Current	Future	%	Δ	%	Δ	%	Δ	%	Δ
1	✓	-	-	-	-	86.5	–	83.2	–	94.2	–	83.9	–
2	✓	MSE	-	-	-	88.6	+2.1	85.9	+2.7	95.3	+1.1	86.8	+2.9
3	✓	Flow	-	-	-	90.1	+3.6	87.8	+4.6	96.2	+2.0	88.5	+4.6
4	✓	Flow	✓	-	-	91.3	+4.8	89.2	+6.0	96.6	+2.4	89.8	+5.9
5	✓	Flow	✓	✓	-	92.3	+5.8	90.4	+7.2	97.1	+2.9	90.9	+7.0
6	✓	Flow	✓	-	✓	92.5	+6.0	90.7	+7.5	97.2	+3.0	91.2	+7.3
7	✓	Flow	✓	✓	✓	**93.2**	**+6.7**	**91.6**	**+8.4**	**97.5**	**+3.3**	**92.0**	**+8.1**

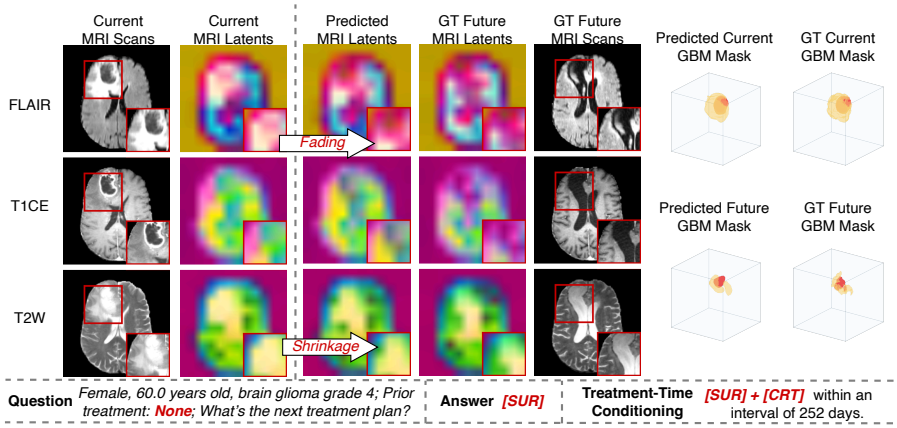


Figure 3: **Qualitative example of treatment-conditioned future prediction.**

Table 4: **Inference latency across standalone treatment planning and on-demand prognostics.**

Execution mode	World-model steps	Latency per case	Relative latency
Treatment planning (Core)	-	**0.375 s**	**1.0×**
+ Temporal mask expert	-	0.563 s	1.5×
+ Latent world expert	50	27.69 s	73.9×

4.5 Efficiency of Inference-Light Treatment Planning

As shown in Table 4, GBM-WAM requires 0.375 s per case for treatment planning. All latency measurements are performed on a single NVIDIA RTX 4090 with a batch size of 1. Enabling the temporal mask expert increases latency to 0.563 s (1.5×), while enabling both auxiliary experts raises it to 27.69 s (≈ 74×), reflecting the additional cost of 50-step Heun integration. Using identical MRI and clinical inputs, the SOTA Brain-WM requires 2.580 s for planning and 41.886 s for generation. Both methods use future-prediction supervision during training and omit future generation during treatment planning. With a compact 0.8B backbone, GBM-WAM achieves higher treatment planning accuracy on both cohorts and approximately 6.9× faster planning than Brain-WM, while retaining optional future-state prediction. Further results appear in Appendix D, while clinical scope and limitations are discussed in Appendix E.

5 Conclusion

GBM-WAM introduces an anatomy-guided latent WAM for training-rich, inference-light glioblastoma treatment planning, using future progression to inform decisions from current MRI and clinical context. Joint world-action modeling shapes treatment-related states via partial-to-complete latent world completion, while temporal mask supervision provides complementary anatomical guidance. Experiments show strong and flexible treatment planning across internal and external cohorts with varying MRI sequences. By retaining supervised query states and bypassing both auxiliary experts, GBM-WAM enables direct treatment planning in one causal MLLM pass. These findings support future prediction as useful training supervision without requiring explicit future generation at inference. Future work will extend this retrospective evaluation through prospective clinical studies and multi-omics integration to assess its value for personalized treatment planning.

Ethics Statement

We acknowledge and adhere to the ICLR Code of Ethics throughout all aspects of our research. This study conducts a retrospective analysis of de-identified MRI scans and clinical records obtained from existing public cohorts under their applicable data use agreements. Ethical approvals, institutional board reviews, and patient consent provisions are fully documented in the corresponding primary publications. No new human participants were recruited for this work.

Reproducibility Statement

Key implementation details, hyperparameter configurations, and evaluation protocols are thoroughly documented in Section 4.1 and Appendix C. Upon publication, all codebase will be publicly released in an open-source repository to facilitate future research.

LLM Usage Statement

We employed the LLM exclusively for language refinement, including grammar correction and stylistic enhancement, to improve readability. All research ideas, methodologies, experiments, analyses, and conclusions are the exclusive work of the authors.

References

Hamed Akbari et al. Machine learning-based prognostic subgrouping of glioblastoma: A multicenter study. *Neuro-Oncol.*, 27(4):1102–1115, 04 2025.

Anthropic. Claude haiku 4.5 system card, October 2025. URL https://www.anthropic.com/claude-haiku-4-5-system-card.

Mahmoud Assran et al. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023.

Shuai Bai et al. Qwen3-VL technical report. arXiv:2511.21631, 2025.

Federico Baldassarre et al. Back to the features: DINO as a foundation for video world models. arXiv:2507.19468, 2025.

Beatrice Bonato, Loris Nanni, and Alessandra Bertoldo. Advancing precision: a comprehensive review of MRI segmentation datasets from BraTS challenges (2012–2025). *Sensors*, 25(6):1838, 2025.

Jun Cen et al. WorldVLA: Towards autoregressive action world model. arXiv:2506.21539, 2025.

Qiuhui Chen and Yi Hong. AliFuse: Aligning and fusing multimodal medical data for computer-aided diagnosis. In *BIBM*, pp. 3082–3087, 2024.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv:2507.06261, 2025.

Tianxingjian Ding, Yuanhao Zou, Chen Chen, Mubarak Shah, and Yu Tian. CLARITY: Medical world model for guiding treatment decisions by simulating context-aware disease trajectories. In *ECCV*, pp. 202–221, 2026.

Daniel Duenias, Brennan Nichyporuk, Tal Arbel, and Tammy Riklin Raviv. HyperFusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. *Med. Image Anal.*, 102, 2025.

Benjamin M. Ellingson et al. Consensus recommendations for a standardized Brain Tumor Imaging Protocol in clinical trials. *Neuro Oncol.*, 17(9):1188–1198, 2015.

Brandon K. K. Fields et al. The University of California San Francisco adult longitudinal post-treatment diffuse glioma MRI dataset. *Radiol. Artif. Intell.*, 6(4), 2024.

Louis Gagnon et al. The University of California San Diego post-treatment glioblastoma (UCSD-PTGBM) annotated multimodal MRI dataset. *Sci. Data*, 13(1), 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, volume 31, pp. 2451–2463, 2018.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Int. MICCAI Brainles. Workshop*, pp. 272–284, 2021.

Edward J. Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Songtao Jiang et al. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. arXiv:2510.08668, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, volume 35, pp. 26565–26577, 2022.

Moo Jin Kim et al. OpenVLA: An open-source vision-language-action model. In *CoRL*, pp. 2679–2713, 2025.

Akhil Kondepudi et al. Health system learning enables generalist neuroimaging models. *Nat. Med.*, 32(8):2831–2837, 2026.

Xunhao Lai, Weiqi Xu, Yufeng Yang, Qiaorui Chen, Yang Xu, Lunbin Zeng, Xiaolong Li, Haohai Sun, Haichao Zhu, Vito Zhang, and Pengyu Zhao. Minimax sparse attention. arXiv:2606.13392, 2026.

LASA Team et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv:2506.07044, 2025.

Yann LeCun. A path towards autonomous machine intelligence. OpenReview, 2022.

Sijie Li, Chen Chen, and Jungong Han. SimMLM: A simple framework for multi-modal learning with missing modality. In *ICCV*, pp. 24068–24077. IEEE, 2025.

Patrick H. Luckett et al. Predicting survival in glioblastoma with multimodal neuroimaging and machine learning. *J. Neuro-Oncol.*, 164(2):309–320, 2023. ISSN 0167-594X.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nat. Commun.*, 15(1):654, 2024.

Esmat Mahmoud et al. MU-Glioma Post: A comprehensive dataset of automated MR multi-sequence segmentation and clinical features. *Sci. Data*, 12(1), 2025.

Mohamed A. Naser and M. Jamal Deen. Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images. *Comput. Biol. Med.*, 121, 2020. Art. no. 103758.

OpenAI. GPT-5.1 Instant and GPT-5.1 Thinking system card addendum. Technical report, OpenAI, 2025. URL https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf.

Maxime Oquab et al. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, pp. 1–31, 2024.

Jiazhen Pan et al. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *MICCAI*, pp. 337–347, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Torsten Rohlfing, Natalie M Zahr, Edith V Sullivan, and Adolf Pfefferbaum. The SRI24 multichannel atlas of normal adult human brain structure. *Hum. Brain Mapp.*, 31(5):798–819, 2010.

Oriane Siméoni et al. DINOv3. *Trans. Mach. Learn. Res.*, 2026.

Yannick Suter et al. The LUMIERE dataset: Longitudinal glioblastoma MRI with expert RANO evaluation. *Sci. Data*, 9(1), 2022.

Chenhui Wang et al. Brain-WM: Brain glioblastoma world model. arXiv:2603.07562, 2026.

Jiguang Wang et al. Clonal evolution of glioblastoma under therapy. *Nat. Genet.*, 48(7):768–776, 2016.

Weiyun Wang et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv:2508.18265, 2025.

Michael Weller et al. EANO guidelines on the diagnosis and treatment of diffuse gliomas of adulthood. *Nat. Rev. Clin. Oncol.*, 18(3):170–186, 2021.

Patrick Y Wen et al. Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. *Neuro Oncol.*, 22(8):1073–1113, 2020.

Patrick Y Wen et al. RANO 2.0: Update to the response assessment in neuro-oncology criteria for high- and low-grade gliomas in adults. *J. Clin. Oncol.*, 41(33):5187–5199, 2023.

Enze Xie et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In *ICLR*, 2025a.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *NeurIPS*, volume 38, 2025b.

Seonghyeon Ye et al. World action models are zero-shot policies. arXiv:2602.15922, 2026.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? arXiv:2603.16666, 2026.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *MICCAI*, pp. 424–433, 2024.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, volume 267, pp. 79115–79135, 2025.

Ying Zhuge et al. Automated glioma grading on conventional MRI images using deep convolutional neural networks. *Med. Phys.*, 47(7):3044–3053, 2020.

Brianna Zitkovich et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *CoRL*, pp. 2165–2183, 2023.

Appendix of Paper: " GBM-WAM: Anatomy-guided Latent World-Action Model for Brain Glioblastoma Treatment Planning"

This appendix supports GBM-WAM with technical details, empirical results, and clinical context. Specifically, it includes: (i) multicenter cohort characteristics and standardized treatment distributions in Appendix A; (ii) clinical and methodological rationales for the five-way discrete action formulation in Appendix B; (iii) implementation details and hyperparameters in Appendix C; (iv) more experimental results, featuring patient-level statistical validation and latent prediction fidelity analyses in Appendix D; and (v) an in-depth discussion on clinical scope, methodological boundaries, and future directions in Appendix E.

A Dataset Characteristics and Treatment Distributions

The source cohorts differ in treatment protocols and selection criteria. For example, LUMIERE (Suter et al., 2022) patients received surgery and temozolomide-based chemoradiotherapy, whereas UCSD-PTGBM (Gagnon et al., 2026) excludes cases with ambiguous treatment effects versus tumor progression. To diversify longitudinal supervision across varying temporal intervals, pairs are constructed densely across all valid visit combinations. The treatment label is assigned to the first treatment recorded after baseline irrespective of the follow-up endpoint, or `[AM]` if none occurred. Thus, pairs sharing the same baseline visit retain the same treatment label, while their auxiliary future-state targets and treatment-time conditions vary with the selected follow-up. We harmonize treatment records into five categories: surgery (`[SUR]`), radiotherapy (`[RT]`), concurrent chemoradiotherapy (`[CRT]`), temozolomide chemotherapy (`[TMZ]`), and active monitoring (`[AM]`, denoting no recorded intervention).

Across the individual source cohorts, the longitudinal pairs reflect complementary facets of clinical care. In LUMIERE ($N = 2{,}112$), the longitudinal pairs are distributed across five labels: routine follow-up without immediate intervention (`[AM]`, $n = 1{,}458$, 69.0%), surgery (`[SUR]`, $n = 363$, 17.2%), concurrent chemoradiotherapy (`[CRT]`, $n = 204$, 9.7%), radiotherapy (`[RT]`, $n = 79$, 3.7%), and standalone chemotherapy (`[TMZ]`, $n = 8$, 0.4%). In contrast, MU-Glioma Post ($N = 842$) provides extensive coverage of ongoing systemic therapy, featuring a substantial share of standalone chemotherapy pairs (`[TMZ]`, $n = 358$, 42.5%) in addition to surveillance (`[AM]`, $n = 373$, 44.3%), concurrent chemoradiotherapy (`[CRT]`, $n = 64$, 7.6%), and radiotherapy (`[RT]`, $n = 47$, 5.6%). Similarly, UCSF-ALPTDG ($N = 298$) is characterized by maintenance and follow-up intervals (`[AM]`, $n = 168$, 56.4%; `[TMZ]`, $n = 112$, 37.6%), complemented by radiotherapy (`[RT]`, $n = 12$, 4.0%) and chemoradiotherapy (`[CRT]`, $n = 6$, 2.0%). UCSD-PTGBM ($N = 83$) contributes an interventional focus, notably capturing recurrent re-resections (`[SUR]`, $n = 28$, 33.7%, predominantly reflecting repeat surgical procedures) alongside observation (`[AM]`, $n = 25$, 30.1%), concurrent chemoradiotherapy (`[CRT]`, $n = 21$, 25.3%), chemotherapy (`[TMZ]`, $n = 5$, 6.0%), and radiotherapy (`[RT]`, $n = 4$, 4.8%).

The evaluation splits capture this combined longitudinal spectrum across diverse settings. The internal test set ($N = 414$) includes all five treatment categories: [AM] ($n = 279$, 67.4%), [TMZ] ($n = 63$, 15.2%), [SUR] ($n = 31$, 7.5%), [CRT] ($n = 31$, 7.5%), and [RT] ($n = 10$, 2.4%). Its category proportions differ from those of the training set, most notably for [AM] (67.4% vs. 60.6%) and [SUR] (7.5% vs. 11.7%). The internal test set ($N = 414$) preserves a proportional composition, providing matched evaluation pairs across `[AM]` ($n = 279$, 67.4%), `[TMZ]` ($n = 63$, 15.2%), `[SUR]` ($n = 31$, 7.5%), `[CRT]` ($n = 31$, 7.5%), and `[RT]` ($n = 10$, 2.4%). The external test set ($N = 83$, sourced from UCSD-PTGBM) serves as an out-of-distribution benchmark enriched with specialized interventional trajectories: `[SUR]` ($n = 28$, 33.7%), `[AM]` ($n = 25$, 30.1%), `[CRT]` ($n = 21$, 25.3%), `[TMZ]` ($n = 5$, 6.0%), and `[RT]` ($n = 4$, 4.8%).

Aggregating diverse clinical cohorts enables the framework to model a wide operational spectrum, bridging standard longitudinal surveillance with specialized interventional trajectories. This curated distribution aligns naturally with real-world neuro-oncology workflows, where observational

monitoring intervals inherently alternate with acute therapeutic decisions. While standardized treatment categories abstract away fine-grained dosing schedules and combinatorial sequencing, this harmonized multi-cohort dataset serves as an initial *methodological proof of concept* to evaluate GBM-WAM for adaptive treatment planning across diverse, heterogeneous clinical settings.

B Rationale for the Five-Way Task Formulation

Real-world clinical planning is complex, multifactorial, and highly heterogeneous, guided by molecular signatures such as MGMT methylation and IDH mutation, performance status, tumor board consensus, and patient preferences. We adopt a simplified five-way classification task for three practical and methodological reasons:

First, longitudinal neuro-oncology cohorts containing paired multi-timepoint MRI and treatment records are scarce, requiring the aggregation of multiple datasets. Harmonizing disparate treatment recording standards into five shared categories provides a consistent prediction target across cohorts. Molecular markers and granular clinical variables are also inconsistently documented or unavailable, shaping the information that can be used consistently for modeling.

Second, the complexity of interactions between disease progression and treatment motivates a tractable methodological proof of concept. A simplified discrete treatment space allows us to investigate whether the world model paradigm can capture treatment-conditioned disease evolution before extending the formulation to more fine-grained regimens.

Third, routinely acquired multi-parametric MRI and standard clinical history provide comparatively accessible baseline inputs across healthcare settings. This input design allows us to investigate the foundational predictive signal available from standard-of-care data and explore its value as a low-barrier preliminary reference for clinicians. Its intended role is to support, rather than replace, comprehensive multidisciplinary evaluation.

C Extended Implementation Details

NeuroVFM (Kondepudi et al., 2026) encodes each observed sequence into a $35 \times 12 \times 12 \times 768$ grid via $4 \times 16 \times 16$ patches, which an adapter projects into the MLLM with sequence and 3D positional embeddings. During training, sequence dropout uniformly samples subsets of one, two, or three visible sequences without replacement, while keeping follow-up targets complete. An 8-head gated cross-attention layer injects factual intervention and interval prompts into future states, allowing auxiliary gradients to update the MLLM while current-mask supervision acts directly on the unconditioned current state. We configure rank-96 LoRA (Hu et al., 2022) with scale 192, dropout 0.05, and gradient checkpointing, with AdamW using $(\beta_1, \beta_2) = (0.9, 0.95)$ and flow matching discretized into 1,000 timestep buckets. All multi-class metrics (F1-score, precision, specificity) are strictly macro-averaged across all five treatment categories for unbiased assessment.

D More Results

Patient-level statistical validation. To explicitly account for intra-subject trajectory correlations and potential sample-size inflation arising from dense longitudinal follow-ups, we evaluated the statistical robustness of GBM-WAM against the strongest baseline Brain-WM. We constructed 95% confidence intervals (CIs) using 1,000 non-parametric patient-clustered bootstrap replicates, where entire longitudinal trajectories of sampled subjects were drawn with replacement. Furthermore, hypothesis testing was conducted via two-sided subject-level paired permutation tests.

On the internal test cohort, GBM-WAM achieved an accuracy of 93.2% (95% CI: [90.8%, 95.4%]) and a F1-score of 91.6% (95% CI: [88.5%, 94.3%]), significantly outperforming Brain-WM at 91.5% (95% CI: [88.6%, 94.0%]) accuracy and 90.4% (95% CI: [87.0%, 93.4%]) F1-score. The paired improvements were statistically significant, yielding a margin of +1.7% (95% CI: [+0.3%, +3.1%], $p = 0.012$) in accuracy and +1.2% (95% CI: [+0.2%, +2.3%], $p = 0.025$) in F1-score.

On the external test cohort, GBM-WAM yielded an accuracy of 81.9% (95% CI: [73.5%, 89.2%]) and a F1-score of 78.2% (95% CI: [69.1%, 86.4%]), surpassing Brain-WM at 78.3% (95% CI:

Table S1: **Cosine similarity across different MRI input subsets on the internal cohort.**

Input Modalities			Modality	Predicted-modality Cosine Similarity (%) ↑			
T2	T1CE	FLAIR	Number	T2	T1CE	FLAIR	Mean
✓				95.26	86.78	85.65	89.23
	✓		1	88.00	94.43	85.39	89.27
		✓		87.95	87.12	94.33	89.80
✓	✓			95.35	94.45	85.30	91.70
✓		✓	2	95.32	86.96	94.33	92.21
	✓	✓		88.04	94.44	94.33	92.27
✓	✓	✓	3	**95.40**	**94.47**	**94.36**	**94.74**

[69.9%, 86.7%]) accuracy and 71.6% (95% CI: [61.1%, 81.3%]) F1-score. The corresponding paired metric gains were +3.6% (95% CI: [+0.6%, +6.8%], $p = 0.018$) for accuracy and +6.6% (95% CI: [+1.9%, +11.4%], $p = 0.005$) for F1-score.

These results confirm that GBM-WAM's clinical planning gains over Brain-WM remain genuine across all cohorts ($p < 0.05$), without repeated-measurement artifacts.

Quantitative results of latent future predictions. To quantitatively assess the prognostic representation quality of the latent world expert, Table S1 reports the cosine similarity between predicted complete future representations $\hat{\boldsymbol{Z}}_{t+1}$ and ground-truth NeuroVFM latents $\boldsymbol{Z}_{t+1}$ across all seven non-empty input subsets. Even when provided with only a single input sequence, GBM-WAM achieves an average cosine similarity of 89.23%–89.80%, demonstrating a meaningful capacity to infer unobserved sequence representations through cross-modal progression priors. As additional imaging modalities become available, the mean cosine similarity increases steadily from 89.23% to 94.73%, with the full three-sequence input yielding the highest overall fidelity across modalities (T2: 95.40%, T1CE: 94.47%, FLAIR: 94.36%). Crucially, evaluating a static identity baseline ($\hat{\boldsymbol{Z}}_{t+1} = \boldsymbol{Z}_t$) yields cosine similarities of 92.94% (T2), 92.74% (T1CE), and 92.43% (FLAIR). Operating in this highly saturated, anatomy-preserving regime, GBM-WAM consistently surpasses this baseline across all modalities, confirming that our MoT-based flow matching captures genuine progression-induced dynamics beyond stationary anatomical redundancy, which directly underpins the downstream planning gains.

E Scope and Limitations

GBM-WAM uses routinely acquired multi-sequence MRI and standard clinical history as a practical, low-barrier foundation for treatment prediction across heterogeneous medical centers. In clinical oncology, molecular signatures (*e.g.* IDH mutation status and MGMT promoter methylation) and patient-specific performance scores (*e.g.* KPS) are critical determinants of multidisciplinary consensus. However, in publicly available longitudinal neuro-oncology cohorts, these markers are frequently undocumented, partially observed, or subject to severe cross-center reporting heterogeneity, precluding their standardized inclusion in unified multi-center benchmarks without discarding the majority of paired longitudinal scans. GBM-WAM is therefore deliberately focused on standard-of-care, routine clinical scenarios where multi-parametric MRI and basic clinical history are universally acquired. Crucially, our choice of an autoregressive MLLM backbone offers inherent architectural extensibility: when such sparse molecular indicators or laboratory values are recorded, they can be natively tokenized as text prompts to enrich the clinical context $c_{\leq t}$ without structural modifications, while its generative vocabulary naturally supports scaling beyond discrete classification toward open-ended treatment regimen formulation.

We also clarify that GBM-WAM focuses strictly on decision-time treatment planning rather than causal counterfactual simulation. Future latent transitions and temporal masks serve purely as auxiliary training-time regularizers over factual progression trajectories to ground intermediate decision states. Because tumor segmentation acts as an auxiliary supervisory prior rather than the primary benchmarked task, optimizing standalone segmentation performance is outside our scope. Quantitatively, the temporal mask expert achieves 71.02% mean Dice on current masks, and 51.20% when forecasting future masks. Delineating fine-grained future voxel-level boundaries from current scans is an intrinsically challenging task due to the pronounced biological heterogeneity and aggressive, non-linear progression of glioblastoma after interventions. We do not pursue high standalone fore-

casting fidelity here; instead, our goal is to leverage these temporal mask signals to provide structural and geometrical guidance during training. As evidenced by our component ablation, incorporating these spatial and longitudinal volumetric change signals consistently enriches the intermediate states and enhances treatment planning performance, while both world and mask experts remain completely bypassed during inference.

We evaluate transfer without external adaptation on an independent cohort of 49 patients and 83 longitudinal pairs. Differences in patient selection and treatment distribution provide a complementary assessment beyond the internal evaluation. GBM-WAM achieves higher accuracy and F1-score than the evaluated baselines on this cohort. Given the limited patient count, uneven treatment-category support, and dependence among pairs from the same patient, these gains provide encouraging evidence of cross-cohort transfer within the evaluated setting, rather than definitive evidence of superiority across broader patient populations.

This study represents an initial step toward clinically grounded world-action modeling. We made extensive efforts to identify, collect, and harmonize all accessible public longitudinal glioblastoma cohorts containing paired multi-timepoint MRI and treatment records. To maximize the utility of these scarce longitudinal observations and capture progression across variable intervals Δt, pairs are constructed densely across valid visit combinations. Crucially, prolonged follow-up intervals inherently heighten decision uncertainty. Our framework mitigates this during training by conditioning future states on factual Δt to absorb temporal variability, whereas inference purposefully bypasses explicit intervals for real-time planning. While pairing a single baseline with multiple follow-up endpoints could theoretically introduce temporal correlation across pairs, our split is strictly executed at the patient level rather than the pair level, preventing data leakage and ensuring independent evaluation on unseen subjects. The remaining benchmark characteristics inherit differences in acquisition protocols, follow-up coverage, and clinical documentation from the source cohorts, which cannot be retrospectively altered. Our efforts therefore focus on making the available data as consistent and usable as possible under unified preprocessing and treatment-label harmonization, providing a foundational baseline for future multi-visit trajectory modeling as larger cohorts emerge.

Looking forward, future work will proceed along four practical directions: *First*, as multi-omics clinical databases mature, we will integrate structured molecular profiles to capture finer-grained biological heterogeneity, alongside exploring spatially masked latent similarity to decouple subregion-specific progression dynamics. *Second*, we aim to leverage the MLLM's autoregressive generation to transition from discrete five-way prediction toward open-ended clinical narratives, including nuanced dosages and combined therapeutic regimens. *Third*, to address the limited size of current public benchmarks, we plan to prospectively collect and validate on larger multi-center cohorts, evaluating GBM-WAM within multidisciplinary tumor boards as a real-time reference. *Finally*, beyond GBM, we will investigate whether anatomy-guided latent world learning transfers to longitudinal disease-progression prediction in other neurological disorders, including mild cognitive impairment conversion, Parkinson's disease progression, and multiple-sclerosis lesion evolution, where incomplete multimodal observations and irregular follow-up pose analogous modeling challenges.

REFERENCES FOR APPENDIX

Louis Gagnon et al. The University of California San Diego post-treatment glioblastoma (UCSD-PTGBM) annotated multimodal MRI dataset. *Sci. Data*, 13(1), 2026.

Edward J. Hu et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Akhil Kondepudi et al. Health system learning enables generalist neuroimaging models. *Nat. Med.*, 32(8):2831–2837, 2026.

Yannick Suter et al. The LUMIERE dataset: Longitudinal glioblastoma MRI with expert RANO evaluation. *Sci. Data*, 9(1), 2022.