Pathology-Controllable Brain MRI Synthesis with Segmentation-Guided Diffusion Models

Anirud Mohan Pranav Shashidhara Tanuka Majumder William Loe
Pramod Kumar

Code Availability. All source code, preprocessing scripts, and evaluation notebooks used in this project are available at: `https://github.com/PranavShashidhara/Seg_diffusion`.

Abstract

Diffusion models have recently shown strong performance in synthesizing realistic medical images, but most existing approaches are uncontrolled. They cannot selectively remove or manipulate pathology while preserving a specific patient's anatomy. This limits their usefulness for explainable AI, data augmentation, and counterfactual reasoning in neuro-oncology. In this project we study pathology-controllable brain MRI synthesis on the RSNA–ASNR–MICCAI BraTS 2021 glioma dataset. Our goal is to generate patient-specific counterfactual FLAIR MRIs. Specifically, when brain tumors are removed while gray matter, white matter, and CSF remain anatomically consistent with the original scan. We build a fully automated preprocessing pipeline that converts 3D BraTS volumes and expert segmentations into tumor-aware 2D slice–mask pairs. On top of this, we implement three diffusion models: (M1) an unconditional baseline, (M2) a segmentation-guided model conditioned on full tissue/tumor masks, and (M3) a Mask-Ablated Training (MAT) variant. For the M3 model, we randomly removes tumor labels during training so the model learns to inpaint healthy tissue in tumor regions. A separate segmentation U-Net (M4) is trained only as an evaluator for our generated images. To evaluate both realism and counterfactual quality, we combine FID, KID, and SSIM with segmentation-based metrics computed by M4: a "GenDice" score for anatomical fidelity of generated images, a residual abnormality ratio (TumorResidual) that penalizes remaining tumor or voids inside the original tumor region, and a Difference-Map IoU that measures how well intensity changes align with the ground-truth tumor. Empirically, M3 achieves the best FID and KID, maintains SSIM and GenDice close to M2, and substantially lowers TumorResidual while slightly improving DiffMap IoU, suggesting that Mask-Ablated Training is an effective and simple mechanism for training diffusion models to generate anatomically plausible, tumor-free counterfactual MRIs.

1 Introduction

Denoising Diffusion Probabilistic Models (DDPMs) and related architectures have achieved state-of-the-art performance in generating realistic medical images. However, standard diffusion pipelines are typically designed for unconditional or weakly conditioned synthesis. These models are abele to sample plausible images from a distribution, but they do not offer fine-grained control over where pathology appears or how it changes. For clinical use, realism alone is not sufficient. Radiologists and researchers often need controllable models that can answer questions of the form: "What would this patient's MRI look like if the tumor were completely removed?"

In brain tumor imaging, high-quality MRI datasets with voxel-wise labels are expensive to collect and are subject to strict privacy regulations. This makes it extremely rare to observe true paired data of the form "tumor-present scan" and "perfectly healthy scan of the same brain." Nevertheless, such pairs would be invaluable: they would enable pixel-level comparison of pathology vs. baseline, support data augmentation for segmentation and prognosis models, and provide a more interpretable way to inspect what an AI system considers abnormal.

1.1 Problem Statement

We study the following problem: given a real FLAIR brain MRI slice and its corresponding tissue/tumor segmentation mask, can we train a generative model that:

1. **Reconstruction:** Reproduces the original pathological scan when conditioned on the full mask (including tumor).
2. **Counterfactual editing:** Produces a tumor-free version of the same slice when the tumor label is removed from the mask, while preserving the patient-specific anatomy of gray matter, white matter, and CSF.

We refer to the second case as patient-specific counterfactual generation, since the output should be a plausible MRI of the same brain under a hypothetical "no tumor" scenario, not a generic healthy brain.

1.2 Motivation and Clinical Relevance

Pathology-controllable counterfactual MRIs are useful in several ways:

- **Explainability and anomaly localization.** By generating a personalized healthy baseline and subtracting it from the original tumor scan, we obtain a difference map that highlights where the model believes abnormalities are. This is more interpretable than a single scalar prediction such as "tumor: yes/no."
- **Synthetic paired data.** Counterfactual pairs (tumor vs. tumor-free) can be used to train and stress-test segmentation and classification models without requiring longitudinal scans or manual construction of healthy images.
- **Simulation and education.** Being able to remove tumors in a controlled way enables scenario-based teaching and exploration of rare or borderline cases for trainees.

These benefits require not just realistic images, but controlled, anatomically consistent edits to pathology.

1.3 Challenges with Existing Approaches

Unconditional diffusion models learn $p(x)$ over images but cannot enforce that a particular region should become healthy. Adding conditioning on segmentation masks is a natural extension, but existing mask-conditioned models are not necessarily trained to respect edited masks. When tumor labels are removed, a naively trained model may:

- ignore the change and still hallucinate tumor-like structure,
- produce black "voids" or blurry artifacts where the tumor used to be, or
- violate anatomical boundaries between gray matter, white matter, and CSF.

Classical inpainting methods also struggle in this setting. Tumor regions are not simple occlusions, and naive interpolation from surrounding pixels often yields unrealistic or clinically meaningless fills. Prior segmentation-guided diffusion works demonstrate the value of mask conditioning, but their behavior under explicit tumor-removal edits in the BraTS setting has not been systematically evaluated.

1.4 Our Approach and Contributions

To address these challenges, we build an end-to-end framework for segmentation-guided, pathology-controllable diffusion on BraTS 2021. Our main contributions are:

1. **Tumor-aware 2D training pipeline.** We design a preprocessing pipeline that converts 3D BraTS volumes into 2D FLAIR slices and four-class masks (GM, WM, CSF, tumor), including intensity normalization and slice selection to focus on informative tumor-bearing slices.

2. **Segmentation-guided diffusion models.** We implement three diffusion models:

   - **M1 (Unconditional):** A baseline DDPM-style model trained only on FLAIR slices.
   - **M2 (Seg-Guided):** A model that concatenates the segmentation mask as an additional channel, learning to generate MRIs consistent with tissue labels, including tumor.
   - **M3 (Seg-Guided + Mask Ablation):** A variant of M2 where, for a subset of training steps, the tumor label is removed from the mask while the image still contains tumor. This forces the model to inpaint plausible healthy tissue.

3. **Counterfactual evaluation protocol with M4.** We train a separate segmentation U-Net (M4) purely as a referee network and use it to define:

   - **GenDice** (Dice score of M4 on generated images) for anatomical fidelity,
   - a **residual abnormality ratio** that penalizes remaining tumor predictions or voids inside the original tumor region in healthy counterfactuals, and
   - a **Difference-Map IoU** that measures alignment between intensity changes and the ground-truth tumor mask.

Using this setup, we empirically compare M1, M2, and M3 and show that Mask-Ablated Training improves the quality of tumor-free counterfactuals relative to a standard segmentation-guided baseline, while maintaining strong realism and anatomical structure.

2 Related Work

Diffusion Models for Medical Imaging. Diffusion models have been widely adopted for medical image synthesis, reconstruction, and cross-modality translation. These models demonstrated strong realism and stable training compared to GANs [6, 7]. Most prior work focuses on learning $p(x)$ or $p(x \mid \text{modality})$ for tasks such as MRI–CT translation, organ synthesis, or accelerated MRI reconstruction, without explicit control over localized pathology.

Recent work on synthetic MR for brain tumor segmentation further emphasizes that standard computer-vision metrics like Inception-based FID and Inception Score can be misleading in the medical domain. For example, Akbar et al. compare GANs and diffusion models for BraTS-style tumor MRI synthesis and propose radiology-aware variants such as Rad-FID and Rad-IS. These variants showcases features that are extracted from a network trained on RadImageNet rather than ImageNet [10]. In line with this and other medical-imaging work, which typically reports FID, SSIM, and segmentation-based Dice/IoU, our metric suite combines global realism (FID) with segmentation-driven fidelity (GenDice) and introduces two tumor-specific measures (TumorResidual and DiffMap IoU) tailored to counterfactual tumor removal.

Segmentation-Guided Generation. Spatial conditioning via segmentation masks or structural priors is a common strategy to enforce anatomical plausibility. GAN-based approaches use segmentation maps to guide lesion placement or organ shape, but often suffer from mode collapse and unstable training. More recent diffusion-based methods, such as SegGuidedDiff [1], concatenate segmentation masks or inject mask embeddings into the U-Net, showing that diffusion models can reliably follow complex tissue layouts. However, these approaches are typically evaluated with fixed, ground-truth masks. Their behavior under deliberately edited masks (e.g., tumor removed) is less explored.

Counterfactual Medical Image Synthesis. Counterfactual medical imaging aims to modify specific pathological regions while leaving the rest of the anatomy unchanged. Prior work has used GANs, VAEs, and early diffusion models to simulate lesion addition or removal, or to study robustness of segmentation networks [2, 8, 9]. These methods often rely on latent space traversals or heuristic inpainting and may sacrifice patient-specific anatomy when performing large edits. In the BraTS glioma setting, there is limited work on truly patient-specific tumor removal with explicit control via segmentation masks.

Mask Ablation and Partial Conditioning. Ideas related to masking and ablation appear in masked autoencoders, partial-label training, and sketch- or scribble-based conditioning, where models learn to infer missing content from context. These approaches suggest that exposing a model to incomplete conditioning at training time can improve robustness at test time. To our knowledge, however, there is no prior work that (i) applies segmentation-guided diffusion to BraTS-like glioma masks, and (ii) explicitly removes tumor labels during training so the model learns to inpaint anatomically plausible healthy tissue in tumor regions.

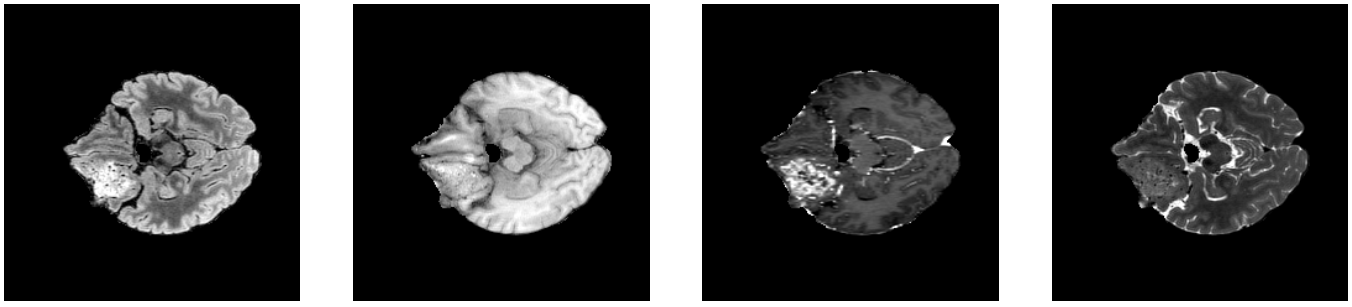
Positioning of Our Work. We build directly on segmentation-guided diffusion and counterfactual medical image synthesis by adapting SegGuidedDiff-style conditioning to BraTS 2021 and introducing Mask-Ablated Training (MAT). Our contributions are: (1) a tumor-aware 2D training pipeline with tissue+tumor masks, (2) a simple MAT scheme that teaches the model to handle edited masks, and (3) an evaluation protocol that directly measures counterfactual quality via residual abnormality and difference-map IoU, rather than relying solely on global realism metrics.

3 Dataset and Preprocessing

3.1 BraTS 2021 Dataset

We use the RSNA–ASNR–MICCAI BraTS 2021 dataset, which contains 2,040 pre-operative glioma MRI studies. Each case includes four co-registered 3D MRI modalities (T1, T1CE, T2, and T2-FLAIR) and expert-annotated segmentation masks for three tumor subregions: necrotic/non-enhancing tumor (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET) (Figure 1). All scans are skull-stripped and co-registered to a common space, which makes them well-suited for slice-wise generative modeling.

In this work we focus on the FLAIR modality, as it provides strong contrast for edema and tumor-related hyperintensities. The original tumor masks are used downstream to define our "pathology" class and to evaluate counterfactual generation quality.



(a) FLAIR (b) T1 (c) T1CE (d) T2

Figure 1: Example BraTS 2021 MRI modalities (T1, T1CE, T2, and T2–FLAIR) for a single case. In this work we focus on T2–FLAIR for generative modeling, while the other modalities are used only for context and potential future extensions.

3.2 Tissue and Tumor Label Construction

To provide richer structural conditioning for the diffusion models, we construct a four-class tissue mask for each slice:

- background / non-brain,
- gray matter (GM),
- white matter (WM),
- tumor (union of NCR/NET, ED, and ET).

Non-tumor tissue labels (GM/WM) are derived via a separate tissue-segmentation pipeline (e.g., an Atropos-style brain tissue segmentation), and then combined with the BraTS tumor mask to form a unified label map. An example of the resulting encoded mask and RGB visualization is shown in Figure 2. This four-class mask is the conditioning signal used by our segmentation-guided diffusion models M2 and M3, and also serves as supervision for the referee U-Net M4.

3.3 Preprocessing Pipeline

We implement a fully automated preprocessing pipeline to convert 3D volumes into 2D slice–mask pairs suitable for diffusion training. The main steps are:

- **Volume-to-slice conversion:** For each case, we extract axial FLAIR slices together with the corresponding 3D tissue/tumor mask, preserving voxel-wise alignment.
- **Slice quality filtering:** Slices with little or no brain tissue (near-empty background) are discarded to avoid wasting capacity on non-informative examples.
- **Spatial normalization:** All slices and masks are resampled to a fixed in-plane resolution using bilinear (image) and nearest-neighbor (mask) interpolation to ensure consistent input size for the diffusion U-Net.
- **Intensity normalization:** FLAIR intensities are clipped to a robust range and linearly rescaled to 8-bit grayscale in $[0, 255]$. During training and sampling, images are further normalized to $[0, 1]$ or $[-1, 1]$ as required by the diffusion pipeline.

- **Mask encoding:** The combined tissue + tumor mask is stored as a single-channel label map with integer values $\{0, 1, 2, 3\}$ (background, GM, WM, tumor). For debugging and visual inspection, we also generate RGB overlays and color-coded mask visualizations.
- **Patient-level splitting and storage:** Train/validation/test splits are performed at the *case* level to avoid leakage across slices from the same subject. Within each split, FLAIR slices and masks are saved with a consistent naming convention `<case_id>_<slice_index>.png`, enabling reproducible dataloading.

This pipeline yields a large 2D dataset of anatomically consistent FLAIR slices paired with four-class tissue/tumor masks. All three generative models (M1, M2, M3) and the referee segmentation model (M4) are trained on this preprocessed representation.

3.4 Preprocessing Outputs

Representative examples of preprocessed FLAIR slices and their corresponding encoded masks are shown in Figure 2, which illustrate the exact inputs seen by our diffusion models and the structural information provided by the conditioning masks. These illustrate the exact inputs seen by our diffusion models and the structural information provided by the conditioning masks. Figure 3 shows the same segmentation mask overlaid on all four MRI modalities, highlighting how pathology appears differently across contrasts.

(a) Input Brain (b) Encoded Mask (c) Original RGB Mask (d) Brain Mask

Figure 2: Example of brain input and corresponding masks used as conditioning for the model. Colors denote background, gray matter, white matter, and tumor (union of NCR/NET, ED, and ET).

3.5 Dataset Split Statistics

Table 1 summarizes the final case-level split and the resulting number of 2D FLAIR slices used for training and evaluation.

4 Methodology

Our methodology is organized around a three–model experimental framework designed to (1) validate the diffusion training pipeline, (2) establish a segmentation–guided baseline which preserves

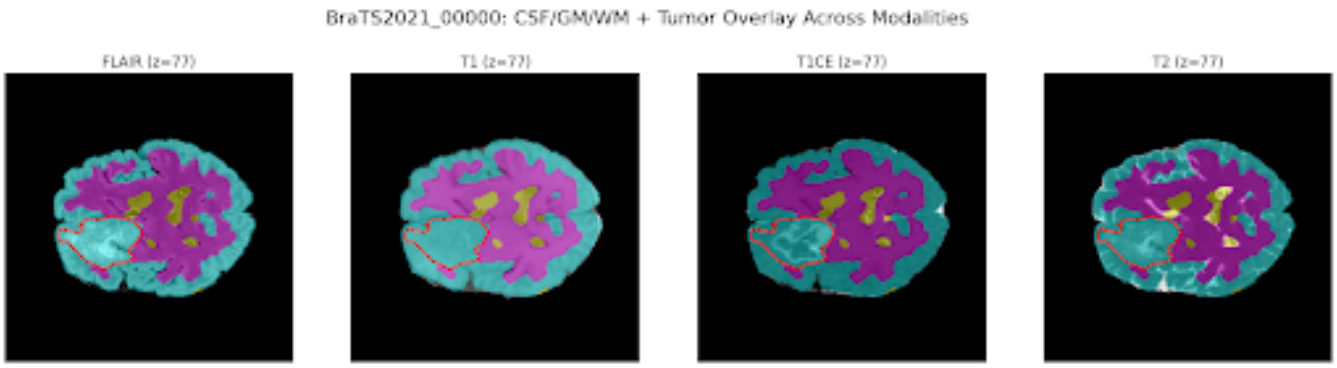


Figure 3: Segmentation Mask Overlayed with different modalities of the brain T1, T2, T1CE and Flair

Split	**# Cases**
Train	80
Validation	10
Test	10

Table 1: Case-level train/validation/test split of BraTS 2021 used in our experiments.

the anatomical accuracy where M1 fails, and (3) test whether Mask–Ablated Training (MAT) improves counterfactual tumor removal while preserving anatomy. All generative models operate on 2D T2–FLAIR slices paired with multi–class tissue/tumor masks derived from BraTS 2021. Throughout this section, M1–M3 denote the generative diffusion models; a separate 2D segmentation U–Net used only for evaluation is denoted **M4** and described later in the Evaluation section.

4.1 Overview of the Model Suite

We train three diffusion–based models:

- **M1 – Unconditional Diffusion Baseline:** A standard DDPM trained only on FLAIR slices, without any mask conditioning. Its purpose is to verify data curation, training stability, and sampling quality.
- **M2 – Segmentation–Guided Diffusion (no MAT):** A conditional DDPM that receives the segmentation mask as an extra input channel. Its purpose is to learn mask–consistent synthesis when conditioning masks are complete thereby preserving anatomical features of the input image.
- **M3 – Segmentation–Guided Diffusion with Mask–Ablated Training (MAT):** Architecturally identical to M2, but with a training–time strategy that randomly removes tumor labels from the conditioning mask in a subset of iterations. Here is where we force the model to learn to inpaint plausible *healthy* tissue in regions where the tumor label is deliberately absent, enabling counterfactual editing.

A high-level overview of the complete preprocessing and three-model pipeline is shown in Figure 7 (Appendix A).

4.2 Diffusion Formulation

All three generators follow the Denoising Diffusion Probabilistic Model (DDPM) framework. Given a clean image x_0, the forward process gradually adds Gaussian noise:

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t;\, \sqrt{\overline{\alpha}_t}\, x_0,\, (1 - \overline{\alpha}_t)\, I\right),$$

where $\{\overline{\alpha}_t\}_{t=1}^{T}$ defines the noise schedule over T timesteps. At training time, the U–Net learns to predict the additive noise ϵ from a noisy image x_t and timestep t.

For the unconditional model (M1), the training loss is the standard DDPM objective:

$$L_{\mathrm{M1}} = E_{x_0,\, t,\, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\, \epsilon,\, t\big)\right\|^2\right].$$

For the segmentation–guided models (M2 and M3), we concatenate a mask m to the noisy image channel and condition the network on both:

$$L_{\mathrm{cond}} = E_{(x_0,m),\, t,\, \epsilon}\left[\left\|\epsilon - \epsilon_\theta\big(\sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\, \epsilon,\, t,\, m'\big)\right\|^2\right],$$

where m' is either the full mask (M2, or M3 in non–ablated steps) or an ablated version of the mask (M3 in MAT steps, described below).

4.3 U–Net Backbone and Conditioning Mechanism

The overall network architecture for M2 and M3 is summarized in Figure 4. All three generators share the same 2D U–Net backbone, closely following the SegGuidedDiff architecture used in our project proposal:

- **Encoder:** The network begins by processing the input image through a series of six residual downsampling blocks. As the spatial resolution decreases, the feature depth progressively increases from 128 up to 512 channels. This hierarchy allows the model to first capture low-level details like tissue textures and boundaries before moving deeper to extract high-level semantic concepts, such as the overall shape of the ventricles or the presence of pathology.
- **Bottleneck:** At the lowest resolution (4×4), the network processes the most compressed representation of the brain using 512 feature channels. Crucially, this stage incorporates a spatial self-attention mechanism that allows the model to look beyond local pixel neighborhoods and integrate global context. This ensures the model understands the overall geometry and relative positioning of anatomical structures before reconstruction begins.
- **Decoder:** The decoder mirrors the encoder, progressively upsampling the features back to the original 256×256 resolution. To prevent the loss of detail common in deep networks, we employ skip connections at every level that copy high-frequency spatial information directly from the encoder to the decoder. This "bridge" ensures that the final generated MRI slices retain sharp edges and anatomical accuracy rather than appearing blurry or over-smoothed.
- **Mask conditioning (M2, M3):** for segmentation–guided models, the input is a two–channel tensor

  $$[x_t \,\|\, m'] \in R^{2\times H\times W},$$

  where x_t is the noisy FLAIR slice and m' is a single–channel label map (scaled to a continuous range) encoding tissue / tumor structure.

M1 uses the same backbone but receives only the noisy image channel x_t as input.

4.4 M1: Unconditional Baseline

The unconditional model M1 serves three roles:

1. **Pipeline sanity check:** confirms that preprocessing, dataloaders, and GPU setup produce stable training.
2. **Realism reference:** establishes a visual and quantitative baseline for sample realism (FID).
3. **Lower–bound anatomy:** provides a reference for how well a model can capture global brain appearance without explicit structural conditioning.

M1 is not expected to support counterfactual editing, but it anchors the performance of the more structured models.

4.5 M2: Segmentation–Guided Diffusion (No MAT)

M2 extends M1 by concatenating the full segmentation mask m to the noisy FLAIR slice at every timestep. In all training iterations, the conditioning mask contains the complete tumor label (class 4) and the healthy tissue classes.

This model learns the mapping

$$\text{(FLAIR, full mask)} \longrightarrow \text{matching MRI slice},$$

and is expected to:

- follow the provided mask geometry (tumor shape and location),
- preserve tissue boundaries between GM / WM / CSF,
- improve FID and Dice–based metrics relative to M1.

However, because it never sees edited masks during training, M2 fails to generate the healthy counterfactual of the input MRI slice at test time.

4.6 M3: Segmentation–Guided Diffusion with Mask–Ablated Training (MAT)

M3 uses the same architecture and hyperparameters as M2 but modifies the conditioning mask during training via Mask–Ablated Training (MAT).

Rather than ablating pixels independently, we follow a simple and clinically motivated rule: with probability p_{MAT} for an entire training iteration, we remove the tumor label everywhere in the mask and treat that region as background:

$$m'(x,y) = \begin{cases} 0, & \text{if } m(x,y) = \text{tumor class (4) and this step is ablative,} \\ m(x,y), & \text{otherwise.} \end{cases}$$

Concretely:

- With probability $(1 - p_{\text{MAT}})$, M3 behaves exactly like M2 and sees the full tumor mask.

- With probability p_{MAT} (e.g., ≈ 0.3 of steps), the tumor region is relabeled as background in m', while the input image x_0 still contains the visible tumor signal.

This creates a deliberate mismatch between the image and the conditioning mask during training. To minimize the denoising error, the model is forced to *inpaint* the tumor region as healthy tissue that is consistent with the surrounding brain. As a result, M3 learns the mapping:

$$\text{(FLAIR with tumor, mask without tumor)} \longrightarrow \text{healthy–looking MRI slice,}$$

which is exactly the counterfactual editing scenario we use at inference time. This MAT behavior is later illustrated qualitatively in Figure 5, where M3 produces visibly healthier counterfactuals than M2 on the same edited masks.

4.7 Training Setup and Hyperparameters

All three models share a common training setup to ensure fair comparison:

- **Input Resolution:** Single–channel T2–FLAIR slices resized to 256×256.
- **Optimizer:** AdamW with a standard learning rate of 1×10^{-4}.
- **Mixed Precision:** We utilized FP16 Automatic Mixed Precision (AMP). This optimization significantly reduced GPU memory consumption and increased training throughput, allowing us to maintain larger batch sizes on the NVIDIA A100.
- **Batch Size:** Kept constant at 64 across all model variants.
- **Diffusion Timesteps:** $T = 1000$ steps for the forward noise schedule during training.
- **Training Duration:** Models were trained for up to 200 epochs. To monitor stability, model weights were checkpointed every 10 epochs, at which point sample images were also generated and logged to WandB for real-time visual assessment of reconstruction quality.

Figure 4: U-Net Architecture Diagram for M2 and M3

By keeping architecture and optimization identical between M2 and M3 and varying only the mask ablation strategy, we can attribute performance differences directly to MAT.

4.8 Inference for Counterfactual Generation

At test time we evaluate both M2 and M3 in two modes:

1. **Reconstruction mode:** We take real FLAIR slice + full segmentation mask (with tumor) as the input. Then we try to reconstruct the original pathological image, checking that structural conditioning behaves as expected.

2. **Counterfactual (healthy) mode:** Here the input is, the real FLAIR slice + edited mask where tumor labels are removed. The goal here is to generate a "healthy" version of the same brain, in which the tumor region is filled with plausible tissue while the rest of the anatomy is preserved.

For both modes we use the same sampling schedule and noise settings across M2 and M3. M1 is sampled unconditionally and used primarily as a baseline for realism (FID) and anatomy quality. Qualitative examples and quantitative metrics (FID, GenDice, residual abnormality ratio, and Difference–Map IoU) are reported in the Results and Discussion sections.

5 Experiments

We evaluated M1–M3 on preprocessed 2D T2–FLAIR slices and tissue/tumor masks from BraTS 2021, focusing on three questions: (i) can a diffusion model learn the BraTS FLAIR distribution, (ii) does segmentation guidance improve realism and anatomy, and (iii) does Mask-Ablated Training (MAT) enable better tumor-removal counterfactuals.

5.1 Dataset Split and Slice Sampling

We constructed disjoint train/validation/test splits at the patient level to avoid leakage across slices from the same subject. From each case, we extracted axial FLAIR slices and:

- discarded near-empty slices with negligible brain tissue,
- ensured that tumor-bearing slices are well represented in all splits,
- resampled images and masks to 256×256 resolution.

All three generative models shared the same train/validation data - differences in performance are therefore, attributable to conditioning and MAT.

5.2 Training Procedure

M1–M3 are trained under identical optimization and diffusion settings:

- AdamW optimizer, learning rate 1×10^{-4};
- batch size 64;
- $T = 1000$ diffusion timesteps during training;
- 1000 denoising steps at inference using the standard DDPM sampler;;

- mean-squared error loss on noise prediction;
- up to 200 epochs with checkpoint selection based on validation loss and sample quality.

M2 and M3 used the same U-Net backbone and hyperparameters; the only difference was the presence of Mask-Ablated Training in M3.

5.3 Counterfactual Inference Protocol

To evaluate counterfactual behavior for M2 and M3, we followed a consistent protocol:

1. Select a test slice with non-empty tumor mask and its ground-truth segmentation.
2. Construct a "healthy" conditioning mask by relabeling all tumor pixels as background.
3. Run the full diffusion sampling process (1000 steps) conditioned on this edited mask to generate a counter-factual image.
4. Compare: (a) M2 vs. M3 counterfactuals on the same slice and mask, and (b) the intensity difference map versus the ground-truth tumor mask.

Unconditional samples from M1 were generated from pure noise and used only as a realism/anatomy baseline. Qualitative examples of reconstruction and healthy counterfactuals for this protocol are shown in Figure 5.

5.4 Evaluation Metrics

We use four metrics on the held-out test set:

- **FID** (Fréchet Inception Distance): distributional realism of generated FLAIR slices (lower is better).
- **GenDice**: Dice score between a referee U-Net's segmentation of generated images and the ground-truth masks (higher is better; anatomical faithfulness).
- **TumorResidual**: residual abnormality ratio inside the original tumor region for "healthy" counterfactuals, based on the referee U-Net and near-void detection (lower is better).
- **DiffMap IoU**: IoU between a thresholded intensity difference map (real vs. healthy image) and the ground-truth tumor mask (higher is better; localized editing).

These metrics jointly assessed realism, structural consistency and the quality of tumor removal.

6 Results

We evaluated our three generators using both qualitative inspection and quantitative metrics targeting: (1) realism, (2) anatomical faithfulness, and (3) counterfactual tumor–removal quality. Throughout this section, we refer to M1 as the unconditional baseline, M2 as segmentation–guided diffusion (no MAT), and M3 as segmentation–guided diffusion with Mask–Ablated Training (MAT). All segmentation–based metrics were computed using the evaluation U–Net (M4).

6.1 Qualitative Results

M1 learned to synthesize brain–like T2–FLAIR slices from pure noise: ventricles, hemispheric symmetry, and smooth tissue gradients were visible, confirming that preprocessing and DDPM training are stable. However, M1 had no notion of tumor geometry or tissue labels, so it could not support controlled editing or counterfactual generation. Representative unconditional samples from M1 are shown in Figure 6, illustrating that the model captures brain-like intensity patterns but lacks anatomical control.

When conditioned on the full tumor mask, M2 produced high–fidelity images that closely followed the mask layout; tumor regions, edema and surrounding tissues aligned well with the conditioning labels. When tumor labels are removed at inference to create a "healthy" mask, M2 produced under–filled or nearly black voids or residual tumor–like texture that broke white/grey matter continuity. This is consistent with its training regime, in which it never sees edited masks.

M3 behaved similarly to M2 in reconstruction mode (full mask), but showed qualitatively different behavior in counterfactual mode: when the tumor label is removed, the former tumor region is filled with tissue that visually matches nearby structures, large voids and bright blobs are avoided, and global anatomy away from the tumor site is better preserved. Side–by–side grids of M2 vs. M3 clearly showed that MAT enabled M3 to "heal" the tumor region in a way that is both visually plausible and anatomically coherent as illustrated in Figure 5.

6.2 Quantitative Evaluation

We reported six metrics on the held–out test set:

- **FID** (Fréchet Inception Distance) and **KID**: global realism of generated FLAIR slices (lower is better).
- **SSIM**: structural similarity between real and generated images in reconstruction mode (higher is better).
- **GenDice**: multiclass Dice between M4's segmentation on generated images and the ground–truth masks (higher is better; anatomical faithfulness).
- **TumorResidual**: residual abnormality ratio inside the original tumor region for "healthy" counterfactuals (lower is better).
- **DiffMap IoU**: IoU between a thresholded difference map (real vs. healthy image) and the ground truth tumor mask (higher is better; measures how well changes localize to the true tumor).

Table 2 summarizes the scores for all three models.

6.3 Interpretation of Metrics

Realism (FID, KID, SSIM): Both segmentation–guided models dramatically improved global realism over the unconditional baseline. M3 achieved the best FID and KID (58.1 and 0.13), followed by M2 (76.0 and 0.21), while M1 was much worse (219.4 and 0.79). SSIM was highest for M2 (0.723) with M3 close behind (0.712), and far above M1 (0.260), indicating that mask conditioning improved structural fidelity.

Model	FID ↓	KID ↓	SSIM ↑	GenDice ↑	TumorResidual ↓	DiffMap IoU ↑
M1	219.403	0.7911	0.2604	0.1541	0.0263	0.0228
M2	75.9503	0.2129	0.7230	0.7136	1.6248	0.1524
M3	58.1257	0.1327	0.7115	0.6881	0.5455	0.1570

Table 2: Quantitative results on the BraTS 2021 test set. M1 is the unconditional diffusion baseline. M2 is the segmentation-guided diffusion model without Mask Ablated Training (MAT). M3 is the segmentation-guided diffusion model with MAT. Lower FID, KID, and TumorResidual and higher SSIM, GenDice, and DiffMap IoU are better.

Anatomical Faithfulness (GenDice): M2 attained the highest GenDice (0.714), showing that M4 can segment its outputs in a way that best matches the ground truth masks. M3 maintained a similar GenDice (0.688), suggesting that MAT does not destroy global tissue structure and trades only a small amount of segmentation alignment for better counterfactual behavior. M1's GenDice was very low (0.154), consistent with its lack of structural conditioning.

Counterfactual Quality (TumorResidual, DiffMap IoU): Tumor–specific metrics are most informative for the counterfactual task. M2 had the worst TumorResidual (1.62), meaning that the original tumor region remained highly abnormal or void–like in its "healthy" outputs, despite its strong reconstruction performance. M3 reduced TumorResidual to 0.55 and achieved the highest DiffMap IoU (0.157), indicating substantially cleaner tumor removal and better localisation of changes to the true lesion area. M1 incidentally had a very low TumorResidual (0.026), but this is misleading: its unconditional samples differed so much from the real image that M4 often did not detect tumor or void in the correct location, and DiffMap IoU remained near zero (0.023). M1, therefore, cannot be considered a meaningful counterfactual model despite its low residual score.

6.4 Overall Findings

Taken together, the results supported our main hypothesis:

- Segmentation guidance (M2/M3 vs. M1) greatly improves realism and anatomical alignment.
- Mask–Ablated Training (M3 vs. M2) is crucial for counterfactual tumor removal: it achieves the best global realism (FID/KID; Table 2), markedly lowers TumorResidual, and slightly improves DiffMap IoU, while maintaining competitive SSIM and GenDice.

In other words, M3 provided the best balance between realism, anatomical structure and clinically meaningful counterfactual edits: M2 is strong for reconstruction but unreliable for tumor removal, and M1 is unsuitable for controllable generation.

6.5 Comparison to Prior Work

Our trends aligned qualitatively with prior segmentation-guided diffusion work. As in SegGuidedDiff [1], adding segmentation masks (M2/M3) dramatically improved segmentation–based realism metrics compared to an unconditional baseline (M1). At the same time, our tumor-specific metrics (TumorResidual and DiffMap IoU) extended earlier evaluations, which typically relied on FID and Dice/IoU alone, by directly quantifying how much tumor signal remains and where image changes

occurred. Following observations by Akbar et al. [10], we treated FID and KID as global but imperfect realism scores and placed additional emphasis on segmentation–driven and tumor–localised metrics when assessing counterfactual quality.

7 Discussion

Our framework evaluates three critical questions: (1) the baseline capability of diffusion on BraTS (M1), (2) the impact of segmentation conditioning (M2), and (3) the necessity of Mask-Ablated Training (MAT) for counterfactuals (M3). Our results demonstrate that while segmentation guidance improves realism, MAT is the essential component that transforms a conditional generator into a controllable counterfactual model.

7.1 The Role of Mask-Ablated Training

The performance gap between M2 and M3 highlights the necessity of ablation. M2, trained only on complete masks, fails to generalize to the "edited" masks used at inference, resulting in artifacts or empty voids where tumors were removed (as seen in the M2 counterfactuals in Figure 5). In contrast, M3 is explicitly trained to handle missing tumor labels, forcing it to inpaint plausible healthy tissue based on surrounding context (bottom row of Figure 5). This is quantitatively confirmed by M3's significantly lower TumorResidual and higher DiffMap IoU scores, proving that ablation enables the model to effectively "heal" pathological regions rather than simply mimicking the mask.

7.2 Balancing Fidelity and Control

While M2 achieves slightly higher GenDice on standard reconstruction tasks, it lacks the flexibility required for editing (Table 2). M3 strikes a superior balance: it maintains high anatomical realism (low FID, high SSIM) comparable to M2, but successfully avoids the "black hole" artifacts seen in M2 counterfactuals (Figure 5). By smoothing out hyperintensities while preserving ventricular shape and cortical boundaries, M3 aligns with findings in recent literature suggesting that explicit lesion handling is required for reliable counterfactual synthesis [1, 8]. Training dynamics for M2 and M3 (validation SSIM and loss over epochs) are shown in Figure 8 and further support that MAT maintains stable optimization while slightly improving reconstruction quality.

7.3 Summary

Ultimately, our experiments confirm that while M1 learns the distribution and M2 enables structural control, only **M3 with Mask-Ablated Training** solves the counterfactual problem. M3 proves to be a simple yet effective method for generating anatomically consistent, tumor-free MRI synthesis, bridging the gap between high-fidelity generation and clinical controllability.

8 Limitations and Future Work

Current Limitations. While the proposed approach demonstrates promising results, several limitations remain. First, all models operate on 2D T2-FLAIR slices, which prevents explicit modeling of cross-slice dependencies and may lead to counterfactual edits that are not fully consistent across

the corresponding 3D volume. Second, the study is restricted to a single imaging modality (T2-FLAIR), whereas clinical decision-making typically relies on the joint interpretation of multiple MRI contrasts, including T1, T1CE, T2, and FLAIR.

In addition, several evaluation metrics used in this work, such as GenDice and TumorResidual, rely on predictions from a referee U-Net segmentation model. As a result, any systematic biases or failure modes in this auxiliary model may propagate into the reported scores. From a modeling perspective, we employ a single U-Net backbone with a limited hyperparameter search due to computational constraints, and diffusion-based sampling (DDPM/DDIM) remains relatively slow when compared to real-time clinical workflows. Finally, this study does not include formal reader studies with neuroradiologists; consequently, the findings should be interpreted as methodological validation rather than direct clinical evidence.

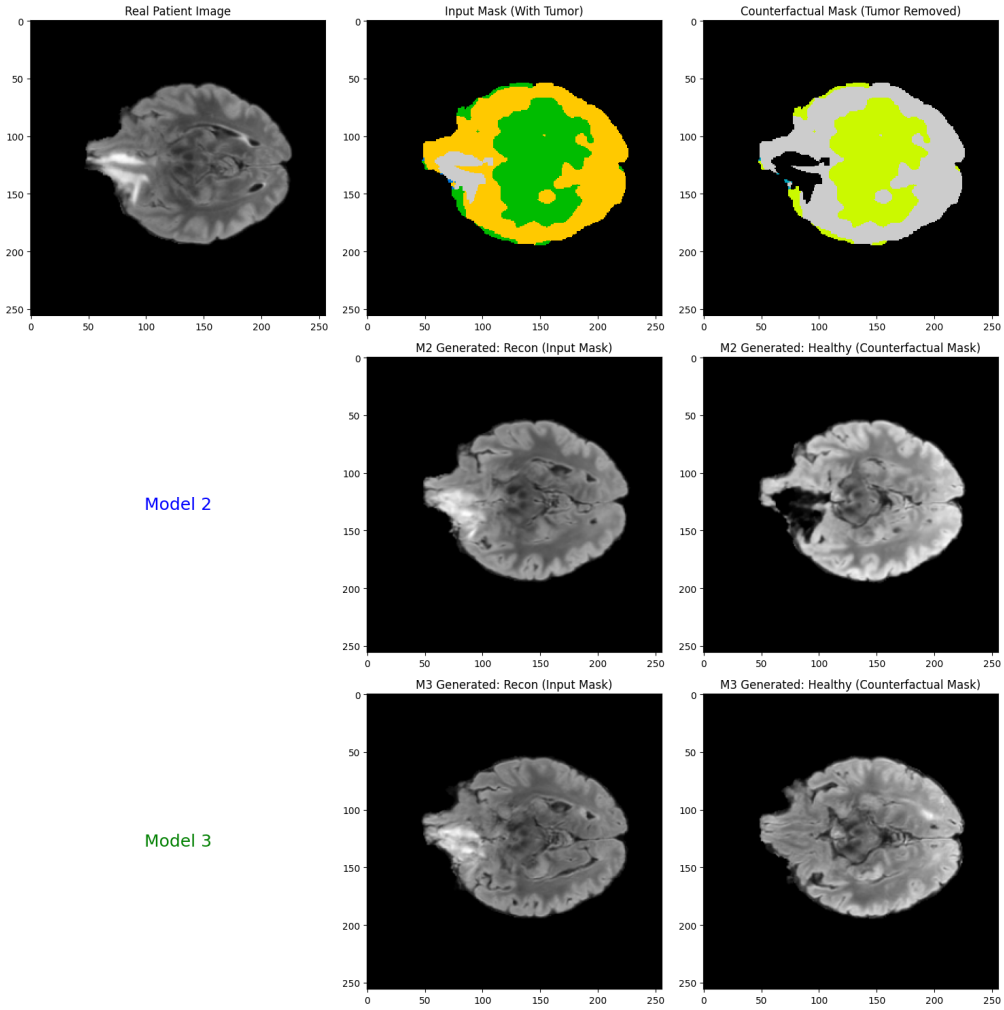


Figure 5: Qualitative comparison of M2 and M3 counterfactuals. Top row: real FLAIR slice, original tumor mask, and edited "healthy" mask. Middle row: M2 reconstruction and healthy outputs. Bottom row: M3 reconstruction and healthy outputs. M3 better inpaints the ablated tumor region with plausible brain tissue.

Future Directions. Several natural extensions follow from these limitations. An important next step is to extend MAT-equipped diffusion models to 3D or 2.5D formulations and to multi-modal

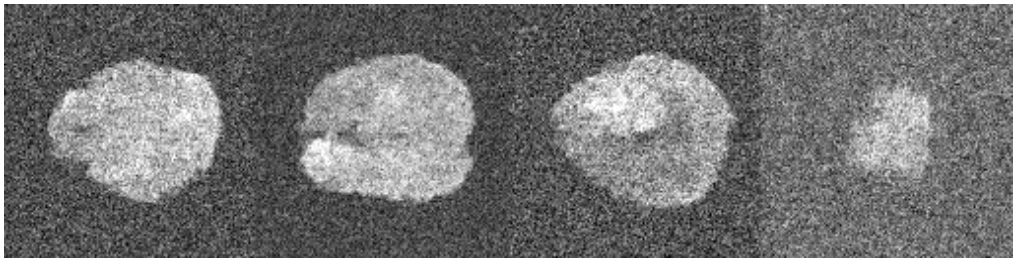


Figure 6: Unconditional samples from M1, illustrating that the model learns brain-like intensity distributions but offers no mechanism for pathology control.

inputs, enabling explicit enforcement of cross-slice consistency while leveraging complementary information across T1, T1CE, T2, and FLAIR contrasts. Beyond lesion removal, richer and more controllable pathology edits—such as tumor resizing, shape perturbations, or treatment-response simulation could be enabled by conditioning on editable masks and higher-level semantic edit descriptors.

On the evaluation side, incorporating classifier-based realism metrics, radiomics features, and task-based performance measures (e.g., the impact of counterfactual images on downstream segmentation or classification models) would provide a more comprehensive assessment of counterfactual validity. These quantitative analyses would ideally be complemented by blinded reader studies and evaluation on external datasets. Finally, the proposed MAT framework is not specific to glioma or MRI; extending segmentation-guided, mask-ablated diffusion models to other organs and imaging modalities, such as liver lesions in CT or lung nodules in CT/PET represents a promising direction for broader methodological and clinical impact.

9 Team Contributions and Implementation Tools

9.1 Team Contributions

All group members collaborated on the literature review, experimental design, and writing. The primary responsibilities were divided as follows:

- **Anirud Mohan**
  Refined dataset curation by integrating healthy tissue labels with tumor masks. Engineered the training pipeline for models M2 (No MAT) and M3 (MAT), implementing Automatic Mixed Precision (AMP) to accelerate training throughput and significantly compress the experimentation timeline. Established robust MLOps protocols, including checkpoint management and granular logging to Weights & Biases with real-time SSIM tracking, and architected an automated inference module for counterfactual evaluation.

- **Pranav Shashidhara**
  Implemented the and created the 5 class segmentation mask using ants and generated 3D volumes for 100 patients with brain mask containing different sections such as Gray matter, White matter, Tumour, cerebrospinal Fluid and background. Contributed to inference the models, generating images for all the models (M1, M2, and M3) and contributing in preparing example figures for the report and slides. Contributed to the methodology and experiments sections of the report. Assisted with generating the inference script for the models.

- **Tanuka Majumder**
  Implemented and trained a 5-class tissue segmentation U-Net used as an external evaluation ("referee") model to assess the anatomical fidelity of generated MRI images. Curated and validated tissue and tumor label mappings, designed and maintained the MRI slice-level preprocessing pipeline (normalization, resizing, empty-slice removal and train/val/test splits), ensured spatial and class consistency across datasets and verified segmentation quality on real and generated images using Dice-based metrics. Led the literature review on diffusion models for medical imaging and counterfactual generation, and authored the related work section.

- **William Loe**
  Implemented the unconditional diffusion baseline (Model M1), including the training loop, scheduler integration, checkpointing, and sampling utilities used to sanity–check the data pipeline and establish a realism baseline. Developed the evaluation pipeline for FID, KID, SSIM, GenDice, TumorResidual, and DiffMap IoU, including slice–level pairing, metric computation, and summary tables. Created the qualitative comparison notebooks and live demo (side–by–side M1/M2/M3 samples and counterfactuals), and coordinated the final presentation and report restructuring.

- **Pramod Kumar**
  Helped refine the experimental protocol (ablation design, counterfactual sampling strategy) and assisted with hyperparameter selection and monitoring of training dynamics using Weights & Biases. Contributed to analyzing failure modes, writing the discussion, and formulating the limitations and future work sections. Supported slide preparation and final editing of the manuscript.

9.2 Implementation Tools and Libraries

All experiments were implemented in Python and conducted primarily on Google Colab, leveraging high-RAM GPU runtimes (e.g., NVIDIA A100) within a Linux environment. This setup enabled efficient training and evaluation of diffusion-based models on high-resolution medical imaging data. The tools and libraries described below were used across different stages of model development, preprocessing, evaluation, and experimentation.

- **Deep learning frameworks:** PyTorch was used for implementing all model architectures and training loops. The HuggingFace diffusers library was employed to support denoising diffusion probabilistic model (DDPM) components, including UNet2DModel, DDPMScheduler, and associated sampling utilities, which formed the backbone of the generative modeling pipeline.

- **Medical imaging and preprocessing:** Volumetric MRI data in NIfTI format were loaded using nibabel. Image registration and tissue segmentation were performed using SimpleITK- and ANTs-based tooling, following an Atropos-style segmentation pipeline. For slice-level processing and visualization, scikit-image and Pillow were used to handle 2D image extraction, normalization, and augmentation.

- **Numerical computing and data handling:** NumPy and SciPy were used for numerical operations, tensor manipulation, and intermediate computations. pandas was used to organize experiment metadata, aggregate evaluation results, and construct summary tables for analysis.

- **Evaluation and visualization:** Model performance was evaluated using custom PyTorch-based implementations, supplemented by existing open-source code where appropriate, to compute metrics such as Fréchet Inception Distance (FID), Dice similarity, and IoU-based measures. Project-specific metrics, including TumorResidual and DiffMap IoU, were used to quantify pathological consistency and structural differences. Matplotlib and Seaborn were used to visualize training dynamics, qualitative sample grids, and difference maps.

- **Experiment tracking and reproducibility:** Weights & Biases (wandb) was used to track training losses, generated samples, and hyperparameter configurations across runs. Version control was maintained using Git and a shared repository, ensuring reproducibility and consistent experimentation across team members.

All source code, preprocessing scripts, and evaluation notebooks are included as part of the project submission. These materials allow for full reproduction of the experiments and results reported in this paper.

10 Conclusion

In this work, we developed and evaluated a segmentation–guided diffusion framework for pathology–controllable brain MRI synthesis, with a focus on generating patient–specific counterfactual images in which glioma tumor regions are removed while preserving underlying anatomy. Our goal was not only to synthesize realistic MRIs, but to do so in a way that supports clinically meaningful "what–if" analyses.

We compared three models: (M1) an unconditional diffusion baseline, (M2) a segmentation–guided diffusion model trained on full masks, and (M3) a segmentation–guided model trained with Mask–Ablated Training (MAT). This design allowed us to isolate the effect of structural conditioning and, crucially, the effect of exposing the model to incomplete masks during training.

Qualitative outputs and quantitative metrics (FID, GenDice, TumorResidual, and Difference–Map IoU) consistently show that M3 outperforms M2 for counterfactual generation: it more reliably removes tumor signal, better preserves non–tumor anatomy, and localizes image changes to the true tumor region. In contrast, M2 follows complete masks well but behaves unpredictably when tumor labels are removed at inference time, underscoring that naive segmentation guidance alone is insufficient for robust counterfactual editing.

These findings support our central hypothesis: teaching a diffusion model to handle ablated or edited masks during training is essential for achieving controllable, anatomically consistent counterfactual MRI synthesis. MAT provides a simple yet effective mechanism for this, and our experiments demonstrate its value in the neuro–oncology setting.

While our approach is currently limited to 2D slices, depends on segmentation quality, and remains computationally intensive, it establishes a concrete path forward. Extending MAT–enabled diffusion models to 3D volumes, multi–modal inputs, and clinically validated evaluation protocols represents a promising direction for future work at the intersection of medical imaging, interpretability, and controllable generative modeling.

11 References

[1] Konz, N., Chen, Y., Dong, H., & Mazurowski, M. A. (2024). *Anatomically–Controllable Medical Image Generation with Segmentation–Guided Diffusion Models*. arXiv preprint arXiv:2402.05210.

[2] Valdenegro–Toro, M. (2024). *Counterfactual Diffusion Models for Medical Image Segmentation and Robustness*. arXiv preprint.

[3] Baid, U., Rane, S., Talbar, S., et al. (2021). *The RSNA–ASNR–MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification*. arXiv preprint arXiv:2107.02314.

[4] Menze, B. H., Jakab, A., Bauer, S., et al. (2015). *The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)*. IEEE Transactions on Medical Imaging, 34(10), 1993–2024.

[5] Bakas, S., Akbari, H., Sotiras, A., et al. (2017). *Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features*. Scientific Data, 4, 170117.

[6] Ho, J., Jain, A., & Abbeel, P. (2020). *Denoising Diffusion Probabilistic Models*. Advances in Neural Information Processing Systems (NeurIPS).

[7] Nichol, A. Q., & Dhariwal, P. (2021). *Improved Denoising Diffusion Probabilistic Models*. Proceedings of the 38th International Conference on Machine Learning (ICML).

[8] Sanchez, P., Kascenas, A., Liu, X., O'Neil, A., & Tsaftaris, S. A. (2022). *What is Healthy? Generative Counterfactual Diffusion for Lesion Localization*. In Medical Image Computing and Computer–Assisted Intervention (MICCAI), BrainLes Workshop.

[9] Pombo, M. E. A., Parisot, S., Schirmer, M. D., et al. (2023). *Equitable Modelling of Brain Imaging by Counterfactual Simulation with Deep Generative Models*. Medical Image Analysis, 84, 102695.

[10] Akbar, M. U., Larsson, M., & Eklund, A. (2023). *Brain tumor segmentation using synthetic MR images: A comparison of GANs and diffusion models*. arXiv preprint arXiv:2306.03414.

A Appendix

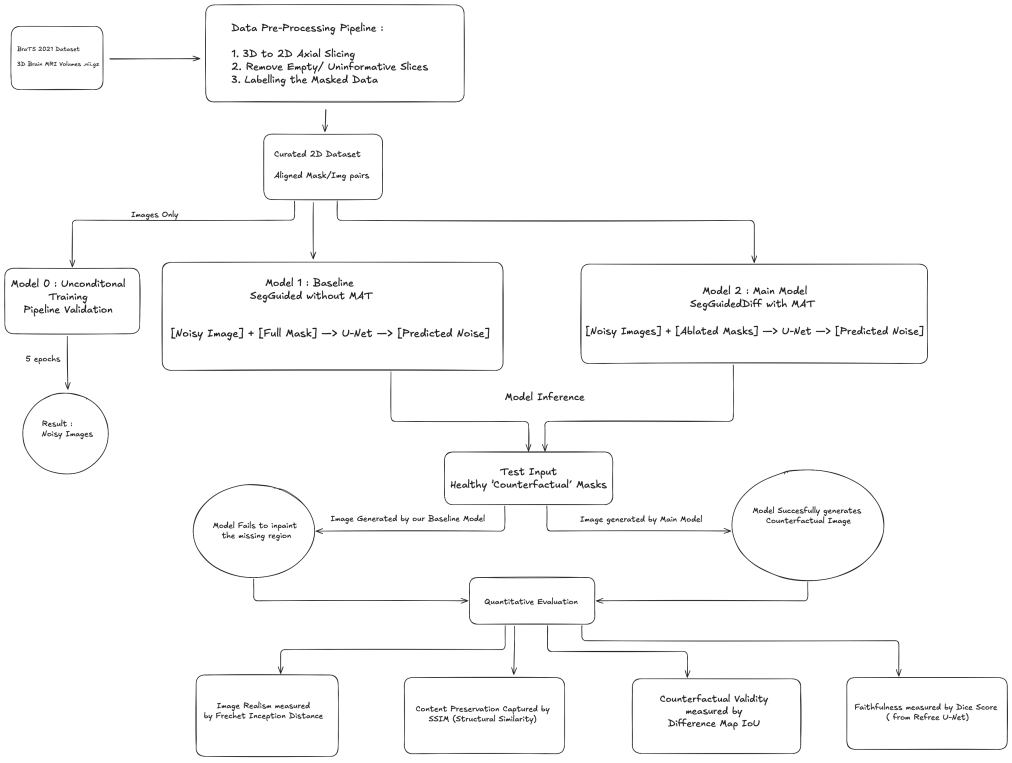


Figure 7: High-level overview of our three-model experimental architecture.

Figure 8: Training Graphs from our WandB