# *GliomaGen*: Unlimited Post-Treatment Glioma MR Images via Conditional Diffusion

Elijah Renner | Thetford Academy

Unless otherwise noted, all images are my own

Project Site

## Abstract

The scarcity of labeled post-treatment glioma MR images limits effective automatic segmentation of key features in brain MR images. Addressing this issue, GliomaGen is introduced, an anatomically informed generative diffusion model that uses a modified Med-DDPM structure to create high-quality MR images from anatomical masks. GliomaGen takes four modalities and six segmentation labels, including a new head area, as input. The developed GliomaGen pipeline augments existing masks to expand the BraTS 2024 Post-Treatment Glioma dataset by 2124 masks, which are later used to synthesize the largest BraTS 2024 Adult Post-Treatment Glioma derivative synthetic dataset (N=2124). Evaluations of GliomaGen with quantitative metrics MS-SSIM, FID, and KID show high fidelity, particularly for t1c (FID: 55.2028 ± 3.7446) and t2w (FID: 54.9974 ± 3.2271) modalities. Segmentation tests with nnU-Net show hybrid training matches real-data performance, but inconsistencies and noise in generated volumes prevented state-of-the-art segmentation from being achieved. These findings show the potential of conditional diffusion models to address data constraints in the BraTS 2024 Adult Post-Treatment Glioma context, and also prompt further iteration on the GliomaGen pipeline.

## Background

### Glioma

Gliomas account for ~80% of malignant brain tumors [1]. Post-treatment MRI glioma analysis is more complex and time-intensive due to changes like resection cavities.

**Dataset:** BraTS 2024 Adult Post-Treatment Glioma $(N = 2200)$. Includes multi-parametric MRI scans (T1, T1-Gd, T2, FLAIR) from seven institutions with four tumor sub-regions (ET, NETC, SNFH, RC).

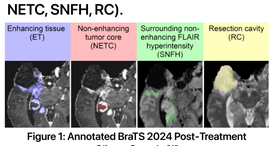



Figure 1: Annotated BraTS 2024 Post-Treatment Glioma Sample [1]

| Institution | Number of Cases (approximate) |
|---|---|
| Duke University | 680 |
| University of California San Francisco | 600 |
| University of Missouri Columbia | 400 |
| University of California San Diego | 350 |
| Heidelberg University Hospital | 300 |
| University of Michigan | 100 |
| Indiana University | 70 |
| **Total** | **2200** |

Figure 2: Institution-Wise Contributions to BraTS 2024 Post-Treatment Glioma [1]

### Data Scarcity Weakens Automatic Glioma Detection

**AI segmentation models** are a promising approach to annotating medical imagery [2], but remain constrained by **data** [3], which **can be difficult to acquire and have annotated by experts.**

### Solution: Synthetic Data using Diffusion

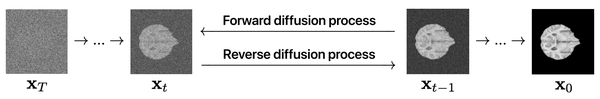



Figure 3: Forward and Reverse Diffusion Processes.

**Diffusion models** have been shown to generate abundant, high-quality synthetic data, as demonstrated by . However,

- Existing diffusion models are not tailored for BraTS 2024 Post-Treatment Glioma segmentation.
- No publicly-available synthetic datasets exist for BraTS 2024 Post-Treatment Glioma.
- Most diffusion pipelines do not use anatomical labels as conditioning, resulting in uncontrollable and outputs and class imbalance.

Towards precise anatomical conditioning, novel frameworks like Med-DDPM [4] and SegGuidedDiff [5] enable diffusion models to use labeled feature masks as instructions, ensuring control. Despite these advancements, these methods have yet to release large annotated synthetic datasets or apply to recent multi-class datasets like BraTS 2024 Adult Glioma.

## Objectives

### Primary Research Question

To what extent can an anatomically conditioned diffusion pipeline generate synthetic MR volumes for the BraTS 2024 Adult Glioma dataset that, when included in training, **improve downstream segmentation performance**, ultimately enhancing diagnostic accuracy and informing better treatment planning for glioma patients?

### Secondary Research Questions

**Segmentation Gains**: How do segmentation models trained with additional synthetic data compare to those trained solely on the existing BraTS 2024 Adult Glioma dataset?

**Scaling the Dataset:** How many synthetic samples are required for meaningful improvements in segmentation performance? What computational resources and trade-offs are involved during generation?

### Hypothesis

By conditioning diffusion models on detailed tumor masks, we can produce high-fidelity, anatomically consistent synthetic MR volumes that **significantly improve glioma segmentation performance** when added to the training set, surpassing current state-of-the-art approaches.

### Engineering Goals

**Leverage Novel Conditional Diffusion Routines:** Implement an anatomically conditioned diffusion model for BraTS 2024 Adult Glioma.

**Large-Scale Synthetic Dataset:** Generate and release the largest public synthetic BraTS 2024 Adult Glioma derivative $(N \approx 1000)$ with multi-class annotations.

**Improvement of Downstream Tasks:** Measure gains in glioma segmentation performance by adding the synthetic dataset into training.

**Clinical Validation Pipeline:** If time allows, collaborate with radiologists to review generated MRI volumes, ensuring anatomical accuracy.

## Methodologies

### [1] Data Preparation

**Head Mask Addition**

0 | Background
1 | New: Head
2| Non-enhancing tumor core
3 Surrounding non-enhancing FLAIR hyperintensity
4 | Enhancing tumor
5 | Resection cavity

**Data Properties**

4 MRI contrasts (t1c, t1n, t2f, and t2w)
Pixel values normalized to $[0, 1]$.
67 subjects removed after mask addition
Resized from $(182, 218, 182) \rightarrow (192, 192, 144)$ for U-Net compatibility.
48 samples set aside for validation, and the final dataset size is $(\text{train, val}) = (1235, 48)$.

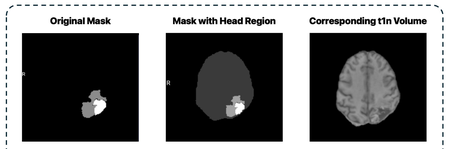
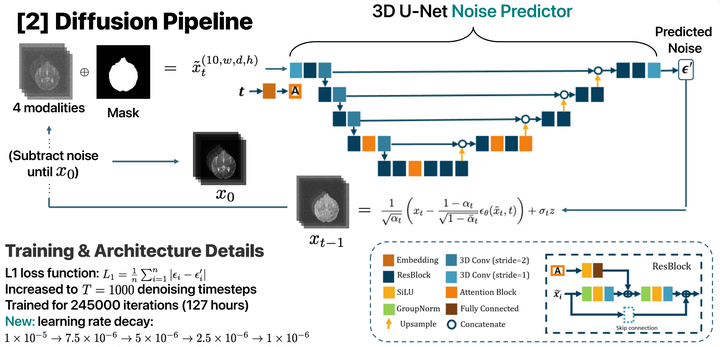



Figure 4: Anatomical Mask Sample Before and After Head Addition [1]

### [2] Diffusion Pipeline




**Training & Architecture Details**

L1 loss function: $L_1 = \frac{1}{n}\sum_{i=1}^{n}|\epsilon_i - \epsilon_i'|$
Increased to $T = 1000$ denoising timesteps
Trained for 245000 iterations (127 hours)
New: learning rate decay:
$1 \times 10^{-5} \rightarrow 7.5 \times 10^{-6} \rightarrow 5 \times 10^{-6} \rightarrow 2.5 \times 10^{-6} \rightarrow 1 \times 10^{-6}$
Cosine noise schedule for controlled noise addition

Figures 5 and 6: Denoising Process Where Added Noise is Predicted by a 3D U-Net. U-Net Architecture Originally Displayed in [4].

### [3] MRI Synthesis

**Mask Augmentation**

Mask $M$

**Mask Transformation Protocol**

1. With probability 0.5, set $M' \leftarrow \texttt{flip}(M)$. Otherwise, $M' \leftarrow M$.
2. Scale $M'$ by a random factor $s$ in the interval $[0.98, 1.02]$ to produce $M''$. [Repeat #2 until $M''$ has a head region within the image bounds].

Figures 7: Mask Augmentation Procedure

- 2 augmentations per mask yielded 2124 transformed anatomical masks
  - ***(346 discarded after failing to preserve the head region)***
- Over 48 hours, GliomaGen was used to generate the corresponding MR modalities for each mask, yielding a synthetic dataset with 2124 subjects.

Modified Mask $M''$

### [4] Data Validation

**Quantitative Metrics**

Used to determine how far generated MRI are from the true MRI for the 48 validation subjects.

**Multi-Scale Structural Similarity Index (MS-SSIM):** checks how similar the generated images are to real MRI scans.

**Fréchet Inception Distance (FID):** Measures distance between Inception-extracted synthetic features and true images.

**Kernel Inception Distance (KID):** Similar to FID but more effective for smaller sample sizes.

**Downstream Task: Segmentation**

To determine the usefulness of synthetic data in downstream tasks, nnU-Net is trained on different data configurations (real, synthetic, high-quality synthetic) to segment real scans.


Figure 8: nnU-Net Segmentation Model Task. Input: 4 MRI contrasts. Output: Labeled Mask.

**Training Data Configurations**

1235R | 1235R + 1062S | 598HQS
1235R + 2124S | 1235R + 598HQS | 617R + 598HQS

**Training Details**

Duration: 1000 epochs
6 stages of 3D convolutions
Kernel size: $[3, 3, 3]$
Batch size: 2
Dice + cross-entropy loss
Learning rate: 0.01 with momentum 0.99 and weight decay $3 \times 10^{-5}$

## Results

### [5] GliomaGen

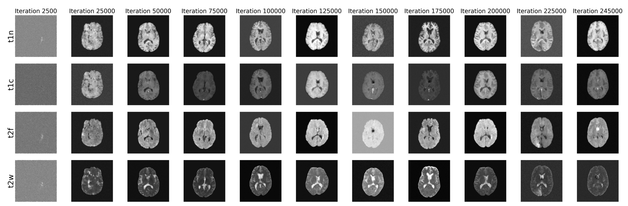



Figure 9: Sample generated modalities over 245000 training iterations

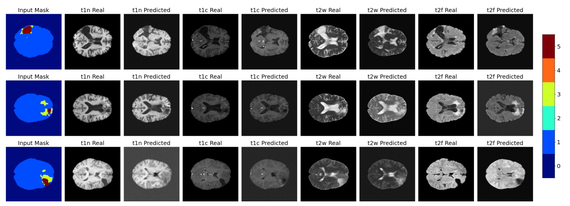



Figure 10: Generated MRI modalities for corresponding validation input anatomical masks with labels non-enhancing tumor core (2), surrounding non-enhancing FLAIR hyperintensity (3), enhancing tumor (4), and resection cavity (5). aside ground truth.

| Modality* | MS-SSIM (↑) | FID (↓) | KID (↓) |
|---|---|---|---|
| t1n[a] | $0.7005 \pm 0.2585$ | $58.4627 \pm 3.8681$ | $0.0305 \pm 0.0011$ |
| t1c[b] | $0.7647 \pm 0.2106$ | $55.2028 \pm 3.7446$ | $0.0293 \pm 0.0019$ |
| t2w[c] | $0.6513 \pm 0.2881$ | $54.9974 \pm 3.2271$ | $0.0291 \pm 0.0010$ |
| t2f[d] | $0.7842 \pm 0.1551$ | $70.4296 \pm 4.1727$ | $0.0370 \pm 0.0018$ |
| **Mean ± SD**† | $0.7252 \pm 0.06$ | $59.7731 \pm 6.5$ | $0.0315 \pm 0.003$ |

Figure 11: Quantitative Results of GliomaGen. Higher MS-SSIM Indicates Better Structural Similarity; Lower FID and KID Suggest Better Perceptual Quality. Values Reported as Mean ± Standard Deviation Across Samples.

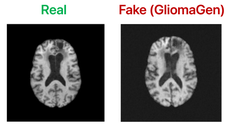



Figure 12: Enlarged Validation t1n Modality vs. Corresponding Synthetic Sample from GliomaGen

**Evaluation**

**Learning:** Figure 9 demonstrates how GliomaGen improves during training.

**Noisy Outputs:** Some generations (e.g., row 3 in Figure 10) exhibited noise and blurriness.

**Mean FID:** 59.7731±6.5; moderately high fidelity with room for improvement.

**Structural Similarity:** MS-SSIM highest for **t2f** and t1c, lower for t1n and t2w.

**Best Modality:** t2w (FID=54.9974±3.2271, KID=0.0291±0.0010).

**Worst Modality:** t2f (FID=70.4296±4.1727, KID=0.0370±0.0018).

### [6] Downstream Segmentation

| Model | Kappa | Accuracy | F1 (NETC) | F1 (FLAIR) | F1 (Enhancing) | F1 (Resection) | Avg Dice | Pearson's r |
|---|---|---|---|---|---|---|---|---|
| Real + 1062 Synthetic | 0.7958 | 0.8969 | 0.7291 | 0.9403 | 0.8915 | 0.9407 | 0.8116 | 0.9919 |
| Real + 598 Synthetic (Subset) | 0.7913 | 0.8945 | 0.7352 | 0.9394 | 0.8898 | 0.9383 | 0.8085 | 0.9938 |
| Real Only | 0.8071 | 0.9033 | 0.7199 | 0.9447 | 0.8951 | 0.9424 | 0.8167 | 0.9923 |
| Real + 2124 Synthetic | 0.7931 | 0.8953 | 0.6998 | 0.9408 | 0.8858 | 0.9363 | 0.7840 | 0.9913 |
| Synthetic Subset Only (598) | 0.4159 | 0.6367 | 0.4457 | 0.7759 | 0.7636 | 0.7024 | 0.5592 | 0.9295 |
| Half of Real (617) + (598) Synthetic (Subset) | 0.7942 | 0.8965 | 0.6962 | 0.9429 | 0.8865 | 0.9316 | 0.8042 | 0.9922 |

Figure 13: Segmentation Performance of nnU-Net Models Trained on Different Data Configurations.

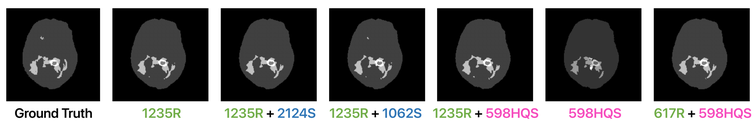



Figure 15: nnU-Net Predicted Segmentation Mask Across Dataset Configurations Aside Radiologist-Annotated Ground Truth

| Model | Avg HD95 (mm) | Avg ASSD (mm) |
|---|---|---|
| Real + 1062 Synthetic | 3.78 | 1.71 |
| Real + 598 Synthetic (Subset) | 4.00 | 1.83 |
| Real Only | 3.35 | 1.43 |
| Real + 2124 Synthetic | 4.01 | 1.70 |
| Synthetic Subset Only (598) | 12.37 | 5.98 |
| Half of Real (617) + (598) Synthetic (Subset) | 4.99 | 1.75 |

Figure 14: Boundary Quality Metrics (HD95 and ASSD) Across nnU-Net Segmentation Models Trained on Different Datasets. HD95 (95th percentile Hausdorff Distance) Measures Worst-Case Boundary error, while ASSD (Average Symmetric Surface Distance) Quantifies Average Boundary Deviation. Lower Values Indicate Better Segmentation Accuracy.

**Key Insights**

**Hybrid data equals real-only performance**
Kappa ~0.79, Accuracy >0.89.

**Synthetic-only underperforms**
Kappa ~0.42, Accuracy ~0.64; worse HD95/ASSD.

**Data efficiency**
Half real + synthetic samples retain strong segmentation.

## Discussion

### Contributions

*Replicable:* open-source pipeline for training GliomaGen and generating synthetic datasets via mask augmentation has applications in other medical domains beyond brain MR images.

*Large dataset:* a publicly-available BraTS derivative dataset is released, serving as the baseline for improvements to GliomaGen.

*Data augmentation:* it was demonstrated that synthetic data has the potential to substantially shrink dataset size while retaining performance.

### Limitations

GliomaGen highlights the effectiveness and pitfalls of applying current diffusion-based methods to complex datasets like BraTS 2024 Adult Post-Treatment Glioma.

Quantitative results indicate that certain MRI modalities, particularly t2f, exhibit elevated noise levels and inconsistent anatomical detail, prompting further evaluation.

Testing revealed that segmentation models trained only on synthetic images did not perform as well as those trained on a mix of real and synthetic images.

*Computational limitations:* testing and revision of methods were constrained by computational costs associated with diffusion, suggesting that efficiency must be improved before widespread adoption.

## Reproducibility

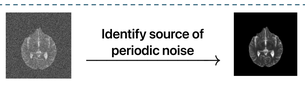
Code available under MIT license on GitHub for immediate adaptation, iteration, and reproduction in the community

GliomaGen weights and BraTS 2024 Adult Post Treatment Glioma-Synthetic dataset provided on HuggingFace.

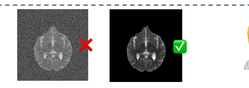
## Future Directions

### GliomaGen Roadmap

**Experimentally refine model architectures and training routines to further reduce inconsistency in generations.**




**Integrate feedback from clinical experts into evaluation.**



**Apply GliomaGen to other BraTS imaging domains**

Sub-Saharan-Africa Meningioma Radiotherapy
Brain Metastases Pediatric Tumors

## Selected References

[1] Maria Correia de Verdier and et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri, 2024. URL https://arxiv.org/abs/2405.18368.

[2] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. Medical Image Analysis, 35:18–31, January 2017. ISSN 1361-8415. doi: 10.1016/j.media.2016.05.004. URL http://dx.doi.org/10.1016/j.media.2016.05.004.

[3] Willemink et al. Preparing medical imaging data for machine learning | radiology. https://pubs.rsna.org/doi/10.1148/radiol.2020192224.

[4] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. IEEE Journal of Biomedical and Health Informatics, 28(7):4084–4093, July 2024. ISSN 2168-2208. doi: 10.1109/jbhi.2024.3385504. URL http://dx.doi.org/10.1109/JBHI.2024.3385504.

[5] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A. Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models, 2024. URL https://arxiv.org/abs/2402.05210.

[6] Muhammad Usman Akbar, M˚ans Larsson, and Anders Eklund. Brain tumor segmentation using synthetic mr images – a comparison of gans and diffusion models, 2024. URL https://arxiv.org/abs/2306.02986.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

## Acknowledgements

- Dr. Alaa Youssef from the AIMI Center for mentorship and guidance
- Dr. Jeremiah Brown from the Geisel School of Medicine for early advice
- Neuroradiologist Dr. Ryan Cusic for discussions about image fidelity
- Vast.ai for providing compute resources