BrainRL: Reinforcement-Driven Computer Vision for Brain mpMRI Analysis of Gliomas

Tim Jing
Department of Computer Science
Stanford University
timjing@stanford.edu

Andrew H. Zhang
Department of Computer Science
Stanford University
andrewhz@stanford.edu

Aaron Lee
Department of Computer Science
Stanford University
aaroncl@stanford.edu

Abstract

Accurate segmentation of gliomas in brain MRI is critical for surgical planning and radiotherapy, yet manual segmentation is time-consuming and subject to variables. AI methods that augment radiologist workflow are key to better patient outcomes. We introduce an end-to-end framework that combines a Deep Q-Network (DQN) agent with a hybrid SAM-UNet segmentation backbone to automate MRI preprocessing (slice selection, windowing, brightness adjustment, and cropping) for improved glioma segmentation. We evaluate our models on the BraTS 2023 dataset (1,251 patients) with an 80/20 train/validation split using over 200 episodes per volume under an ε-greedy schedule (decaying to 0.05). The resulting RL-SAM-UNet achieved an F1/Dice score of 0.9008 versus 0.8535 for static SAM-UNet and 0.7894 for U-Net, demonstrating a modest improvement in the more challenging cases. Our RL-SAM-UNet framework not only achieves state-of-the-art accuracy on the BraTS dataset but also operates with inference times compatible with real-time clinical use, demonstrating the viability of integrating RL agents with vision models to augment physician workflows.

1 Introduction

1.1 Clinical Need

Gliomas are among the most common brain tumors, and high-grade glioblastomas (a type of glioma) are extremely aggressive with poor prognosis even with treatment (6.9% five-year survival rate) Glioblastoma Foundation (2021). Approximately 480,000 people are diagnosed with gliomas annually Mesfin et al. (2024). Accurate identification of tumor tissue in MRIs is crucial for clinical decision-making, including surgical planning, radiotherapy targeting, and monitoring treatment. Expert radiologists can perform this task, but it is often time-consuming and difficult Bakas et al. (2017). Any AI-enabled segmentation methods that can increase efficiency or accuracy would dramatically improve patient care.

1.2 Research Gap

While there has been extensive research in medical segmentation, we believe there is a notable research gap. First, modern computational pipelines for medical segmentation rely on siloed algorithms.

Stanford CS224R 2025 Final Report

Research is often entirely rooted in computer vision, with small variations in model architecture leading to marginal improvements in metrics. Within architectures, researchers often stay in narrow boxes, adjusting various CNN-based or transformer-based architectures. This is not how a physician works in a hospital setting. Radiologists are expected to conduct slice selection, pre-processing, contrast selection, and more manually before providing their analysis. Vision models are unable to conduct these important tasks that are essential for successful integration in a real-time hospital setting.

1.3 Project Objective and Summary

To address the research gap, we aim to develop an RL agent that can learn effective MRI preprocessing (windowing, contrast, slice selection, etc.) to improve diagnostic clarity and achieve SOTA vision model segmentation on a dataset of gliomas. To our knowledge, this has not been done before. To benchmark against this objective, we aim to evaluate two gold standard baseline vision models. The first is U-Net Ronneberger et al. (2015), a go-to vision model for medical segmentation. The second is a novel SAM-UNet encoder-decoder model that we develop ourselves, which we later show achieves state-of-the-art segmentation by combining transformer and convolutional features. We aim for our RL agent to outperform both baselines, even despite the novelty of the SAM-UNet architecture.

2 Related Work

2.1 Work On Vision Models

There has been significant research in vision models for brain tumor segmentation due to the aforementioned clinical need. Historically, a number of algorithms including clustering, edge/contour-based methods, and others have been explored Yao et al. (2023). Subsequently, researchers learned that algorithms based on CNNs (Convolutional Neural Networks) empirically perform well. Over the past few years, the gold standard CNN method for medical segmentation has been UNet Ronneberger et al. (2015), an encoder-decoder model introduced in 2015. The UNet has been adapted to handle 3D images as well in the VNet architecture Milletari et al. (2016). However, CNNs struggle with long-range dependencies, since convolutions depend on local features.

Recently, the introduction of attention-based methods like Vision Transformers (ViTs) has promised to revolutionize medical segmentation Dosovitskiy et al. (2021). Attention enables long-range dependencies to be modeled, theoretically providing more relevant clinical information. MedSAM 2 has achieved high accuracy on a broad range of segmentation tasks as currently serves as the gold standard. The generalization arises from its Masked Autoencoder pretraining on a large dataset of unlabeled images Kaiming He (2021). Subsequent user prompts are decoded into segmentation masks by the decoder Zhu et al. (2024); Kirillov et al. (2023).

Thus, for our baselines, we select the U-Net to represent convolutional networks and develop a novel ViT-CNN model, SAM-UNet.

2.2 Limitations in Vision and Prior Work In RL

Despite the advancements of both CNN and transformer-based methods, the models still lack an adaptive mechanism to handle scan-specific variations. A key limitation of such static models is their inability to generalize. For instance, a network trained on one hospital's scanner may underperform on another without fine-tuning (Stember and Shalu (2022)). This has prompted interest in methods that adapt to each image, which is precisely the niche that we intend to fill using RL. RL has shown promise in computer vision tasks that require sequential decision-making or attention control. Mnih et al. introduced an RL-based attention model that learns to focus on relevant image regions sequentially, rather than processing an entire image uniformly (Mnih et al. (2014)). In object detection for natural images, Caicedo et al. demonstrated that an agent could learn to localize objects by iteratively adjusting a bounding box via deep Q-learning (Caicedo and Lazebnik (2015)). These works highlight how RL vision approaches can outperform passive, single-pass models by intelligently selecting regions of interest. Building on these ideas, researchers have started applying RL to medical imaging. Ghesu et al. developed one of the first RL agents for medical images, which learned to navigate 3D CT volumes to find anatomical landmarks (Ghesu et al. (2016)). Subsequent studies extended this to pathology localization. Maicas et al. used a DQN agent to actively detect breast lesions in

MRI, moving a 3D window to zoom in on tumors (Maicas et al. (2017)). These pioneering works demonstrated that RL can handle the complex task of medical image segmentation. However, they also revealed limitations. Prior approaches usually fix certain preprocessing steps (such as filtering or lung region cropping) beforehand.

Our project addresses these gaps by combining an RL agent with a vision model and explicitly allowing the agent to control preprocessing parameters. In doing so, we aim to handle more complex scenarios (like multiple lesions across dozens of MRI slices) and improve upon the rigidity of earlier methods.

3 Method

3.1 Data Source Description

We use the BraTS 2023 multimodal brain tumor MRI dataset Baid and et al. (2021) Menze et al. (2015). This dataset contains pre-operative MRI scans from multiple institutions. Each patient underwent a multi-parametric MRI (mpMRI) scan including T1-weighted, post-contrast T1-weighted (T1Gd), T2-weighted, and T2-FLAIR imaging. Expert radiologists manually annotated the scans with ground truth segmentations for the tumor subregions. A set of example images for a patient is displayed in Figure 1.

The three labels correspond to: (1) the enhancing tumor core, (2) the edema surrounding the core, and (3) the necrotic tumor core. Each patient contains the four MRI scans listed above with a ground truth segmentation mask. The training dataset is 1251 patients, with $240 \times 240 \times 155$ voxel images. The patient cohort ranges from high-grade glioblastomas to low-grade gliomas. The test set is 219 patients Menze et al. (2015) Baid and et al. (2021). We focus on the training set for model development since ground truth data is not available for the test set. Splits are detailed in the experimental setup section below.

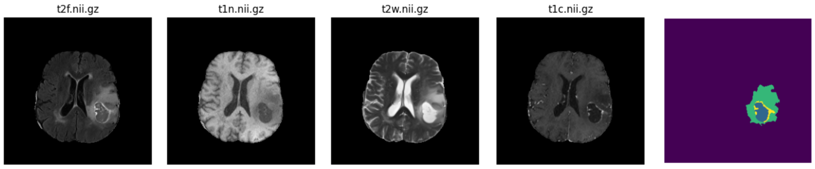


Figure 1: The four provided MRI images with the ground truth segmentation mask on the right for a single patient.

3.2 Methods Overview

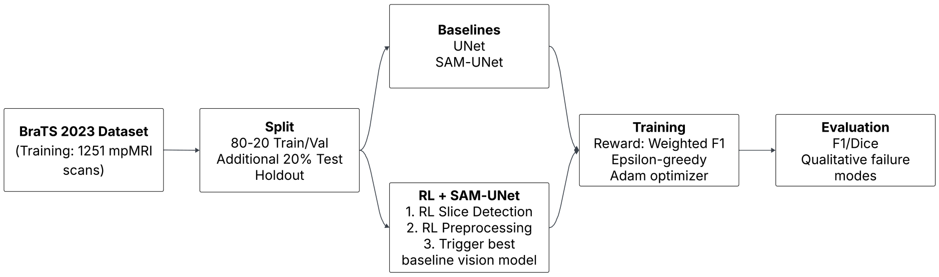


Figure 2: Overall BrainRL methods flowchart.

3.3 U-Net

We apply the standard U-Net architecture Ronneberger et al. (2015). It comprises of 4 layers of encoder/decoder DoubleConv blocks, with a final 1x1 convolution to output the four multi-class segmentation labels.

3.4 SAM-UNet

We implement a hybrid ViT-UNet architecture using a frozen SAM ViT encoder and a custom U-Net decoder. The U-Net decoder incorporates modules that fuse the ViT features through different upsampling convolution scales, which allow for both fine-grained and high-level structures to be captured. The model directly outputs multi-class probabilities (background, NCR, ED, ET) as opposed to U-Net's traditional binary classification. By incorporating a U-Net, we capture relevant spatial information that may be harder to capture in a transformer decoder attending to all patches.

3.5 RL Agent (Deep Q-Network + SAM-UNet)

Task Formulation: Our environment consists of three-dimensional brain MRI volumes, where the agent's state at each time step comprises the current two-dimensional slice image. The image is augmented by any applied brightness and contrast adjustments and the region of interest crop as well as metadata including the slice index, brightness level, and contrast window. To navigate this environment, the agent's action space enables dynamic control over image preprocessing and spatial focus: it can adjust the contrast window or brightness, expand or shrink the crop region, move one slice forward or backward through the volume, trigger a SAM-UNet segmentation inference on the current crop, or terminate the episode to output a final segmentation mask.

RL Agent and Reward: We employ a Deep Q-Network (DQN) that takes as input the processed slice image I_t together with the metadata vector $m_t = (\text{slice\_idx}, \text{brightness}, \text{contrast})$. A DQN was selected due to the discrete action space. Exploration follows an ε-greedy policy, with ε decaying to 0.05 over the course of training. Each episode is capped at 40 steps, and we conduct 200 episodes per volume in the training phase. Upon termination—whether by the agent's explicit terminate action or by reaching the maximum step count—the final crop is passed through the SAM-UNet segmentation model, and the terminal reward r_{term} is calculated as the weighted Dice similarity coefficient between the predicted mask P and the ground truth mask G, scaled to lie within $[-1, +1]$:

$$r_{\text{term}} = 2\frac{|P \cap G|}{|P| + |G|}.$$

While we initially introduced a small negative penalty during each non-terminal step to encourage concise action sequences, we removed it in hopes of encouraging more aggressive actions taken by the RL model. Additionally, whenever the agent triggers an intermediate SAM-UNet inference mid-episode, we compare that inference's Dice score to the best score achieved so far and issue a positive reward for improvements or a penalty for any regression. This combination of terminal, stepwise, and intermediate rewards balances the twin objectives of high segmentation accuracy and efficient exploration.

To assess the impact of network architecture on performance, we implemented and evaluated two DQN variants: a baseline DQN (v1) with three standard convolutional layers and a more sophisticated enhanced DQN (v2). The DQN (v2) consists of deeper convolutional layers with residual connections to facilitate gradient flow, a spatial attention mechanism to focus on salient image regions, improved metadata fusion through additional fully connected layers, and an optional dueling architecture that separately estimates values and advantages before combining them. These enhancements were designed to improve the agent's ability to extract meaningful visual features from MRI slices and make more informed decisions about crop positioning and segmentation timing.

DQN (v1): Three convolutional layers followed by fully connected layers:

$$\text{Conv layers:} \quad \text{Conv2d}(1 \rightarrow 32, k = 8, s = 4) \rightarrow \tag{1}$$

$$\text{Conv2d}(32 \rightarrow 64, k = 4, s = 2) \rightarrow \tag{2}$$

$$\text{Conv2d}(64 \rightarrow 64, k = 3, s = 1) \tag{3}$$

$$\text{Image features:} \quad \phi_{\text{img}}(I_t) = \text{FC}(\text{flatten}(\text{conv\_output}), 512 \rightarrow 256) \tag{4}$$

$$\text{Metadata features:} \quad \phi_{\text{meta}}(m_t) = \text{FC}(m_t, 3 \rightarrow 64 \rightarrow 32) \tag{5}$$

$$\text{Q-values:} \quad Q(s_t, a) = \text{FC}([\phi_{\text{img}}(I_t); \phi_{\text{meta}}(m_t)], 288 \rightarrow 128 \rightarrow 13) \tag{6}$$

DQN (v2): Deeper architecture with residual connections, spatial attention, and dueling structure:

$$\begin{aligned}
\text{Backbone:}\quad & \text{Conv2d}(1 \to 64, k = 7, s = 2) \to && (7)\\
& \text{MaxPool} \to && (8)\\
& \text{ResBlocks}(64 \to 128 \to 256 \to 512) && (9)\\
\text{Attention:}\quad & \text{SpatialAttn}(x) = x \odot \sigma(\text{Conv2d}(x, 512 \to 1)) && (10)\\
\text{Dueling:}\quad & Q(s, a) = V(s) + A(s, a) - \frac{1}{|\mathcal{A}|}\sum_{a'} A(s, a') && (11)
\end{aligned}$$

RL Training: Training follows the DQN algorithm with experience replay and target networks. The agent interacts with MRI volumes for episodes of maximum 40 steps, using ε-greedy exploration with ε decaying from 1.0 to 0.05 over training. The replay buffer stores transitions $(s_t, a_t, s_{t+1}, r_t, \text{done}_t)$.

Q-learning uses temporal difference update rule where θ^- represents the target network parameters:

$$Q_\theta(s_t, a_t) \leftarrow Q_\theta(s_t, a_t) + \alpha \left[r_t + \gamma \max_{a'} Q_{\theta^-}(s_{t+1}, a') - Q_\theta(s_t, a_t)\right] \tag{12}$$

Integration with SAM-UNet: Our segmentation backbone is the SAM-UNet architecture fine-tuned on brain MRI tumor segmentation using the BraTS dataset. Within each episode, the agent decides when to invoke the SAM-UNet model on the current cropped view. The resulting mask both informs the agent's immediate reward (through intermediate comparisons) and determines the terminal reward. By coupling the DQN policy with the high-capacity SAM-UNet model, the agent learns to present optimally informative views that maximize segmentation performance while minimizing unnecessary actions.

Evaluation: We compare our approach against the standalone SAM-UNet baseline using the standard segmentation metric: Dice score, or F1 score. We also measure the average number of actions per episode to assess efficiency improvements over exhaustive or random search. Finally, we conduct a qualitative analysis of the agent's action trajectories to verify that learned strategies, such as initial coarse scanning followed by focused zooming on suspected tumor regions, align with intuitive radiologist workflows.

4 Experimental Setup

All experiments were conducted on the BraTS 2023 dataset with an 80/20 patient-wise train/validation split. We performed hyperparameter tuning covering learning rates, discount factor, batch sizes, replay buffer capacities, and ε decay schedules using grid search. The target network was synchronized every 100 learning steps. Episodes were limited to a maximum of 40 actions. All implementations were built in PyTorch, with final model selection based on the highest validation Dice coefficient.

Table 1 includes a comprehensive list of hyperparameters for our final RL model.

5 Results

Figures 3, 4, 5 highlight quantitative training curves for U-Net, SAM-UNet and RL + SAM-UNet respectively.

U-Net, Figure 3: The left panel shows both training and validation loss steadily decreasing over 30 epochs. The two curves remain closely aligned throughout, indicating minimal overfitting. The right panel plots the corresponding Dice similarity coefficient. Validation Dice is slightly inferior to training.

SAM-UNet, Figure 4: The left plot reports F1 scores by tumor label: the Background class quickly reaches near-perfect performance, while the Enhancing Tumor and Edema classes perform admirably. Necrotic Core is the most challenging segmentation class. The right plot illustrates the SAM-UNet's training loss over 100 epochs, declining smoothly. These trajectories confirm that the model learns easy distinctions first and then progressively refines minority-class predictions.

RL + SAM-UNet, Figure 5: In the upper left panel, the agent quickly navigates from slice 77 to slice 79 by step 3 and remains there for focused analysis. The upper right panel shows brightness and

Hyperparameter	**Value**
Training Parameters	
Learning Rate	5×10^{-5}
Optimizer	Adam
Batch Size	32
Replay Buffer Size	100,000
Target Network Update Frequency	100 steps
RL Parameters	
Discount Factor (γ)	0.99
Epsilon Start	1.0
Epsilon End	0.05
Epsilon Decay	80% of total episodes
Reward Structure	
Dice Reward Weight	10.0
Improvement Bonus (Intermediate Rewards)	2.0
Action Space	
Number of Actions	13
Brightness/Contrast Step	1.0
Crop Movement Step	8 pixels

Table 1: Training hyperparameters for the DQN-based RL preprocessing agent.

contrast adjustments. The lower left panel tracks the crop's IoU with the ground-truth mask, rising to 0.11 once the final inference is triggered at step 18. Finally, the Gantt chart (lower right) details each action in sequence—slice moves, windowing changes, cropping operations—culminating in the inference action. Together, these plots demonstrate that the learned policy prioritizes rapid slice localization followed by iterative windowing and cropping refinements to maximize segmentation overlap.

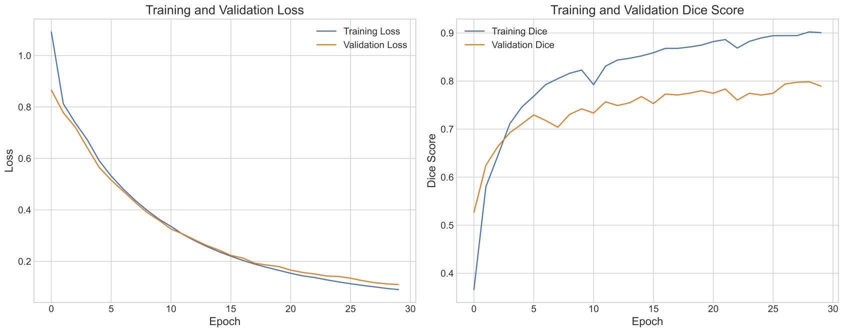
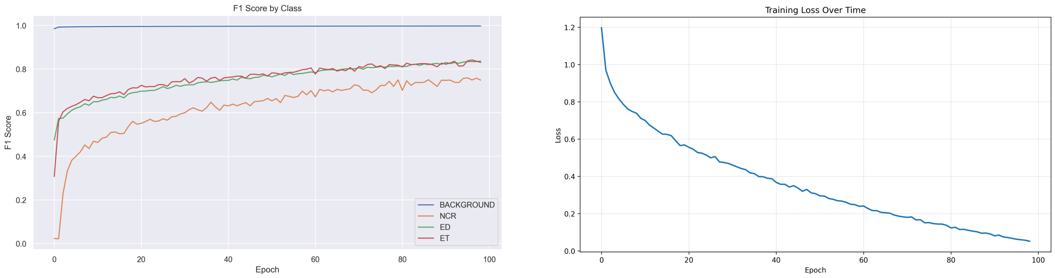


Figure 3: Quantitative results for U-Net training.



(a) F1 Score by Class

(b) ViT-UNet Dice + CE loss

Figure 4: Quantitative results for SAM-UNet training.

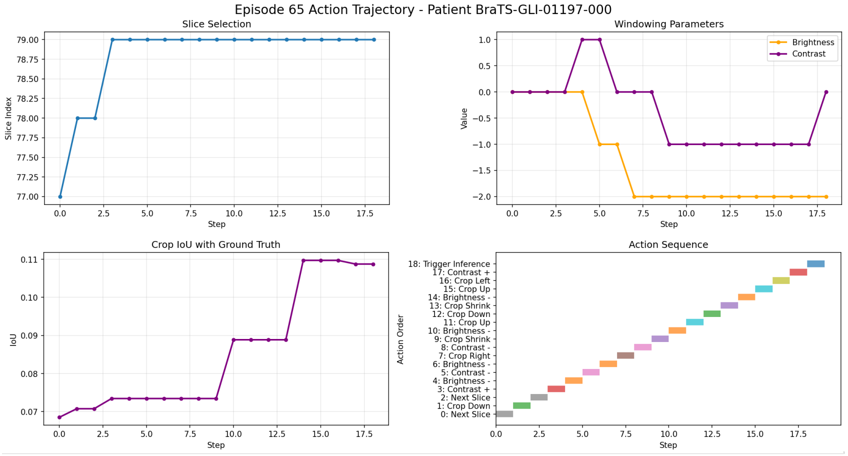


Figure 5: Quantitative trajectory for RL model.

5.1 Quantitative Evaluation

Table 2 shows the performance of our three models on the BraTS dataset in F1 segmentation overlap. The standard U-Net baseline attains a respectable Dice of 0.7894, reflecting its already strong capability for segmentation. The SAM-UNet model yields a substantial improvement to 0.8535.

Introducing the RL agent atop SAM-UNet further boosts performance to 0.8943. This +0.0408 gain over SAM-UNet demonstrates that guiding the segmentation network with intelligent slice selection, windowing, and cropping enables the model to focus on segmenting the tumor.

Furthermore, the enhanced DQN (v2) architecture achieved a mean F1 score of 0.9008, representing a +0.0065 improvement over the initial RL-SAM-UNet (v1) implementation, indicating some additional performance benefit by improving the agent's ability to extract meaningful features.

Table 3 breaks down SAM-UNet performances across individual tumor subregions. While SAM-UNet demonstrates strong performance on edema (F1 = 0.8364) and enhancing core (F1 = 0.8306) segmentation, the necrotic core, which is more homogeneous with surrounding tissue and has irregular boundaries, sees comparatively lower accuracy (F1 = 0.7494). We hypothesize that the RL-SAM-UNet's increased average Dice score partially stems from the use of windowing and contrast adjustments to better elucidate more diffuse boundaries characteristic of the necrotic core, enabling more precise delineation of this challenging subregion.

Overall, these results confirm that combining multi-slice navigation and dynamic preprocessing control leads to state-of-the-art segmentation accuracy on challenging glioma MRI data.

Table 2: Comparison of segmentation methods on BraTS 2023 (F1 scores)

Method	**U-Net**	**SAM-UNet**	**RL-SAM-UNet (v1)**	**RL-SAM-UNet (v2)**
Mean F1 Score	0.7894	0.8535	0.8943	0.9008

Table 3: SAM-UNet segmentation metrics by class.

Metric	**Background**	**Enhancing Core**	**Edema**	**Necrotic Core**
F1 Score	0.9975	0.8306	0.8364	0.7494

5.2 Qualitative Analysis

Figure 6 presents a representative overview of qualitative segmentation masks from all three methods. While U-Net and SAM-UNet performed slightly worse, the failure modes were consistent and represented in the figure. Additional qualitative images are included in the appendix.

While the main tumor mass is successfully identified, it exhibits several characteristic failure modes. First, the predicted mask (bottom-left panel) shows random scattered predictions in the healthy brain, reflecting a tendency for the agent to predict high-contrast regions as tumors. This is something we aimed to alleviate with more aggressive cropping, except it seems the agent did not learn to emulate this behavior. We discuss methods to alleviate this failure mode below in Future Directions. Second, the boundary between enhancing tumor and necrotic core is blurred in the prediction (bottom-right), indicating that the model sometimes confuses these adjacent subregions. This is problematic because the necrotic core is the most difficult class, and thus more effective contrast that does not apply to the whole image might help ameliorate the issue. Third, small tumors at the periphery of the primary mass are occasionally missed. Finally, adjustments in brightness and contrast across the entire scan can introduce artificial intensity gradients in healthy-presenting regions, which lead to undersegmentation of low-contrast necrotic core regions. Please refer to the appendix for other examples of similar failure modes.

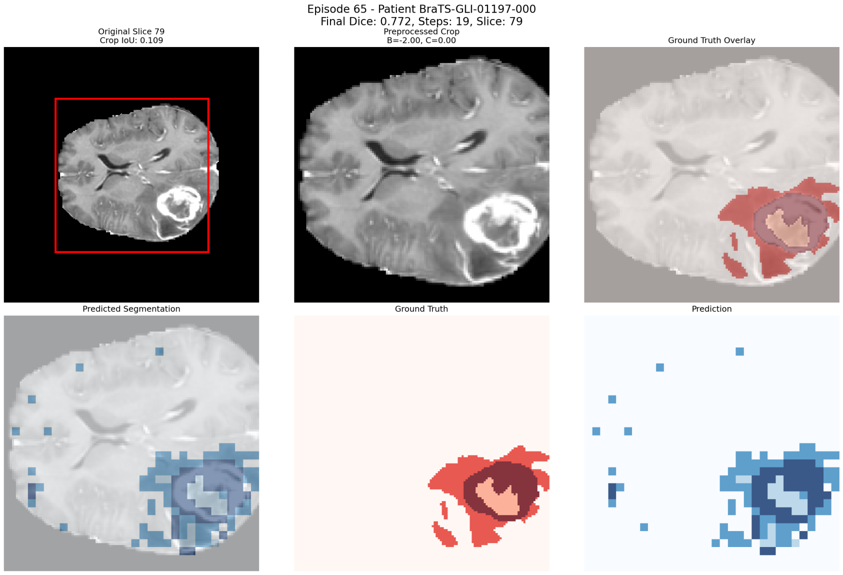


Figure 6: RL-SAM-UNet qualitative segmentation.

6 Discussion

6.1 Interpretation

Our experiments demonstrate that augmenting a SOTA SAM-UNet segmentation backbone with an RL–based preprocessing and slice-selection yields measurable improvements in tumor delineation performance. The RL-SAM-UNet pipeline achieved F1 scores of 0.8943 (DQN v1) and 0.9008 (DQN v2)—surpassing both the standard U-Net (0.7894) and the SAM-UNet baseline (0.8535)—indicating that adaptive windowing, cropping, and navigation policies can effectively complement attention-driven architectures. These gains, though incremental, were most pronounced in challenging MRIs where fixed preprocessing could not reveal tumor boundaries.

6.2 Limitations

Despite these improvements, several limitations constrain the current approach. First, with regard to the problem formulation, the RL-SAM-UNet system remains limited to single-slice processing due to computational constraints, potentially neglecting full three-dimensional context and spatial continuity, where the agent never observes how tumor morphology evolves across neighboring slices in the full MRI. Second, the RL agent's F1 gains may be partially inflated by the vision model's fine-tuning on the same dataset, suggesting that true generalization across unseen institutions requires further validation.

Third, the ϵ-greedy exploration strategy may not fully capture the diversity of optimal preprocessing policies, and abrupt brightness or contrast adjustments can introduce artifacts that mislead the segmentation network. It is possible that in a 13-action state space, the only dense feedback comes from intermediate Dice improvements, which could trap the policy in a local, not global, optima. Finally, the overall performance improvement is modest relative to SAM-UNet, implying that RL optimization may yield diminishing returns when applied to already high-performing architectures.

7 Conclusion

7.1 Future Directions

Looking ahead, several avenues could enhance the RL-SAM-UNet paradigm. Encouraging more aggressive cropping through modifying our reward to include Ground Truth Crop IoU, incorporating clinical utility metrics (e.g., reducing physician review time or prioritizing high-risk regions) into the reward function, and training an upstream agent to decide when segmentation is necessary could further tailor the system for real-world workflows. We also hope to add targeted contrast and image processing options such that the model does not have to edit the entire image each time. Together, in this hierarchical abstraction, basic actions could also be grouped together so the top-level agent would only need to select some macro-action from a reduced action space as opposed to 13 singular edits every time.

Semi-automated pipelines where the RL agent flags low-confidence or outlier cases for expert review might make the model more clearly relevant. Obtaining radiologist feedback on a visualization of the agent's action sequence during training could similarly aid in refining clinically interpretable outputs, similar to an RLHF model. This may also give way to distill a validated policy into a lightweight, single-pass network for even faster inference.

Lastly, evaluation on a broader range of diverse datasets with adversarial features would confirm robustness and generalizability.

7.2 Achievements and Clinical Implications

This work establishes that reinforcement learning can serve as an effective addition to modern segmentation networks, capturing gains in performance by dynamically controlling the data preprocessing. The RL-SAM-UNet framework achieves SOTA segmentation on a challenging dataset with minimal inference time. This directly corresponds with our goal of real-time augmentation of radiologist workflows. Thus, we hope our model with future improvement could be a candidate for hospital deployment, improving diagnostic consistency, accelerating case throughput, and improving patient outcomes.

8 Team Contributions

- **Tim:** Initial proposal research, literature review, U-Net baseline, SAM-UNet baseline, RL agent development and experiments, milestone report/final report/final presentation writing.
- **Andrew:** Supported proposal writing, RL agent formulation and development, RL agent experiments, presentation/final report writing.
- **Aaron:** RL agent experiments, milestone report background research/drafting/experiment, poster presentation writing, final report writing.

Changes from Proposal For contributions, we added Aaron as a team member and thus distributed contributions accordingly. Additionally, while Tim was originally the point person for RL and Andrew was the point person for Vision, we swapped these roles due to Tim's stronger domain expertise with vision segmentation models. In addition, we changed from a pulmonary CT segmentation task to a more well-defined glioma MRI segmentation task. "General lung tumor segmentation" wasn't specific or actionable, and we found this glioma dataset with a very similar task with strong clinical relevance. No other substantial changes were made.

References

U. Baid and et al. 2021. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv:2107.02314.

S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, and et al. 2017. Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Nature Scientific Data, 4:170117. https://doi.org/10.1038/sdata.2017.117

Juan C. Caicedo and Svetlana Lazebnik. 2015. Active Object Localization with Deep Reinforcement Learning. arXiv:1511.06015 [cs.CV] https://arxiv.org/abs/1511.06015

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929 [cs.CV] https://arxiv.org/abs/2010.11929

Florin C. Ghesu, Bogdan Georgescu, Tommaso Mansi, Dominik Neumann, Joachim Hornegger, and Dorin Comaniciu. 2016. An Artificial Agent for Anatomical Landmark Detection in Medical Images. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (Lecture Notes in Computer Science, Vol. 9902)*, Sebastien Ourselin, Leo Joskowicz, Mert R. Sabuncu, Gozde Unal, and William Wells (Eds.). Springer, Cham, 229–237. https://doi.org/10.1007/978-3-319-46726-9_27

Glioblastoma Foundation. 2021. Glioblastoma Survival Rate: A Comprehensive Guide for Patients and Loved Ones. https://glioblastomafoundation.org/news/glioblastoma-multiforme. Accessed: 2025-05-29.

Saining Xie Yanghao Li Piotr Dollár Ross Girshick Kaiming He, Xinlei Chen. 2021. Masked Autoencoders Are Scalable Vision Learners. arXiv:2111.06377 [cs.CV] https://arxiv.org/abs/2111.06377

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. arXiv:2304.02643 [cs.CV] https://arxiv.org/abs/2304.02643

Gabriel Maicas, Gustavo Carneiro, Andrew P. Bradley, Juan E. Nascimento, and Ian Reid. 2017. Deep reinforcement learning for active breast lesion detection from DCE-MRI. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer, 665–673.

B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, and et al. 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). , 1993–2024 pages. https://doi.org/10.1109/TMI.2014.2377694

F. B. Mesfin, T. Karsonovich, and M. A. Al-Dhahir. 2024. *Gliomas*. StatPearls Publishing, Treasure Island (FL). https://www.ncbi.nlm.nih.gov/books/NBK441874/ Accessed: 2025-05-29.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. 2016. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. arXiv:1606.04797 [cs.CV] https://arxiv.org/abs/1606.04797

Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. Recurrent Models of Visual Attention. arXiv:1406.6247 [cs.LG] https://arxiv.org/abs/1406.6247

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs.CV] https://arxiv.org/abs/1505.04597

J. N. Stember and H. Shalu. 2022. Reinforcement learning using Deep Q networks and Q learning accurately localizes brain tumors on MRI with very small training sets. *BMC Medical Imaging* 22 (2022), 224. https://doi.org/10.1186/s12880-022-00919-x

Wenjian Yao, Jiajun Bai, Wei Liao, Yuheng Chen, Mengjuan Liu, and Yao Xie. 2023. From CNN to Transformer: A Review of Medical Image Segmentation Models. arXiv:2308.05305 [eess.IV] https://arxiv.org/abs/2308.05305

J. Zhu, A. Hamdi, Y. Qi, Y. Jin, and J. Wu. 2024. Medical SAM 2: Segment medical images as video via Segment Anything Model 2. arXiv preprint arXiv:2408.00874 [cs.CV]. https://arxiv.org/abs/2408.00874

A Additional Figures

Figure 7 and Figure 8 are additional qualitative segmentation masks.

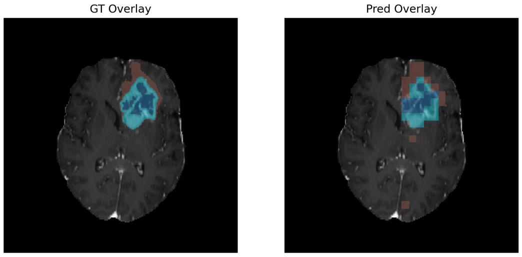


Figure 7: SAM-UNet qualitative segmentation. Masks were of lower resolution and upsampled.

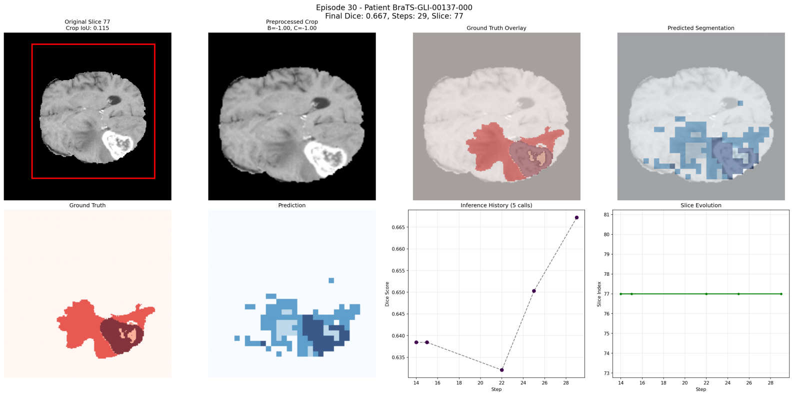


Figure 8: Another qualitative segmentation from the RL agent. Failure modes are consistent with the described report.