CoSeg-MARL: COmmunicative Multi-Agent RL for brain tumor SEGmentation

Anonymous authors
Paper under double-blind review

Abstract

Accurate segmentation of brain tumors from multimodal magnetic resonance imaging (MRI) remains challenging due to the heterogeneous appearance of tumor sub-regions, including necrosis (NEC), edema (ED), and enhancing tumor (ET), and the need to maintain consistency among anatomically related segmentation predictions. In this paper, we propose CoSeg-MARL, a communicative multi-agent reinforcement learning framework based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) for iterative brain tumor segmentation from multimodal MRI. The framework employs three specialized agents that collaboratively refine segmentation predictions through cooperative sequential decision-making. Each agent specializes in one of three hierarchically related tumor regions whole tumor (WT), tumor core (TC), and enhancing tumor (ET) and coordinates with the other agents through shared observations, shared convolutional representations, and explicit message passing. Each agent generates continuous refinement actions comprising four global segmentation controls and a coarse spatial coefficient grid, which are transformed into updated segmentation masks through a deterministic applicator. A hybrid training objective combines supervised Dice and binary cross-entropy losses with a reinforcement learning policy-gradient term, while the reward function encourages segmentation improvement and penalizes hierarchical inconsistencies. Experiments on 2D MRI slices from a patient-level subset of BraTS 2020 evaluate the proposed framework under two initialization strategies: refinement from zero-initialized logits (cold start) and refinement from pretrained U-Net predictions (warm start). CoSeg-MARL achieves mean Dice scores of 0.821 and 0.916, respectively. Compared with the pretrained backbone, warm-start refinement improves the Dice scores for whole tumor, tumor core, and enhancing tumor segmentation from 0.811, 0.830, and 0.838 to 0.927, 0.916, and 0.907, respectively.

1 Introduction

Brain tumor segmentation is an important task in neuro-oncology, particularly for gliomas, which exhibit considerable heterogeneity in their biological characteristics, progression, and clinical outcomes (Rasool & Bhat, 2025). Accurate delineation of tumor regions from medical images supports diagnosis, treatment planning, and longitudinal disease monitoring. Gliomas vary in their histological and molecular characteristics and are classified according to the World Health Organization (WHO) classification of central nervous system tumors (Louis et al., 2021). Their heterogeneous appearance and infiltrative growth patterns make accurate identification of tumor boundaries and sub-regions challenging. Reliable segmentation of these regions is therefore essential for quantitative tumor assessment and clinical decision-making.

Magnetic resonance imaging (MRI) is widely used for brain tumor assessment because of its ability to provide complementary soft-tissue contrast across different imaging sequences. Multimodal MRI, including T1-weighted, contrast-enhanced T1-weighted (T1c), T2-weighted, and fluid-attenuated inversion recovery (FLAIR) images, provides complementary information for identifying tumor boundaries and characterizing heterogeneous tissue regions (Bakas et al., 2017). The BraTS benchmark defines distinct tumor components, including necrotic and non-enhancing tumor core (NCR/NET), peritumoral edema (ED), and enhancing tumor (ET) (Menze et al., 2014). These components are combined into three clinically relevant segmentation regions: whole tumor (WT), tumor

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

core (TC), and enhancing tumor (ET), which follow the hierarchical relationship $\mathrm{ET} \subseteq \mathrm{TC} \subseteq \mathrm{WT}$. However, substantial variations in tumor location, size, shape, and appearance, together with overlapping intensity distributions between tumor and surrounding healthy tissue, make accurate segmentation challenging (Bakas et al., 2018). In addition to delineating individual tumor regions, segmentation methods must account for their anatomical relationships to produce hierarchically consistent predictions.

Recent advances in deep learning have substantially improved medical image segmentation. Convolutional neural networks, particularly encoder-decoder architectures such as U-Net (Ronneberger et al., 2015) and self-configuring frameworks such as nnU-Net (Isensee et al., 2021), have established effective approaches for learning spatial and contextual representations from medical images. Despite these advances, segmenting heterogeneous brain tumor regions remains challenging, particularly when tumor boundaries are ambiguous and predictions for anatomically related regions must remain consistent. Conventional feed-forward segmentation networks typically generate predictions in a single inference pass, although iterative refinement can also be incorporated into supervised architectures. This motivates investigating sequential refinement strategies in which segmentation predictions are progressively updated while accounting for the relationships among tumor regions.

Deep reinforcement learning (DRL) provides a framework for learning sequential decision-making policies, allowing segmentation to be formulated as a Markov Decision Process (MDP) in which an agent iteratively modifies its predictions according to a reward signal (Alansary et al., 2019). Unlike single-pass prediction, this formulation allows the refinement process to account for previous segmentation states and the consequences of successive actions. However, brain tumor segmentation involves multiple anatomically related regions whose predictions cannot be treated as entirely independent. For instance, an enhancing tumor prediction must remain within the tumor core, which in turn must remain within the whole tumor. This relationship motivates a cooperative multi-agent reinforcement learning (MARL) formulation in which specialized agents refine individual tumor regions while exchanging information to coordinate their decisions. Such a formulation provides an opportunity to combine region-specific refinement with shared contextual information and hierarchical consistency constraints.

In this paper, we propose CoSeg-MARL, a communicative multi-agent reinforcement learning framework based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) (Lowe et al., 2017) for iterative brain tumor segmentation from 2D slices of multimodal MRI. The framework comprises two stages. First, a 2D U-Net backbone is trained to generate initial segmentation logits, providing a pretrained initialization for the subsequent refinement stage. The backbone is then frozen, and three specialized agents are trained to refine predictions for whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The agents coordinate through shared observations, shared convolutional representations, and explicit message passing. Each agent predicts continuous refinement actions comprising four global segmentation controls and a coarse spatial coefficient grid, which are converted into updated segmentation predictions through a deterministic applicator. A hybrid training objective combines supervised Dice and binary cross-entropy losses with a reinforcement learning policy-gradient term, while the reward encourages improvements in segmentation quality and penalizes hierarchical inconsistencies and unnecessary refinement steps. We investigate both refinement from zero-initialized logits (cold start) and refinement from pretrained U-Net predictions (warm start) on the BraTS 2020 dataset (Menze et al., 2014). The key contributions of this paper are as follows:

- **Cooperative multi-agent formulation for hierarchical tumor segmentation:** We formulate brain tumor segmentation as a cooperative sequential refinement problem and propose CoSeg-MARL, a communicative MADDPG-based framework with three agents specialized in the hierarchically nested WT, TC, and ET regions. The agents coordinate through shared observations, shared convolutional representations, and explicit message passing.

- **Communicative actor architecture with continuous segmentation refinement:** We design a U-Net-inspired, agent-conditioned actor that integrates multimodal MRI features with current segmentation masks and logits. The architecture combines shared feature extraction, agent-specific conditioning, and communication to produce continuous refinement actions comprising four bounded global controls and a coarse spatial coefficient grid. A deterministic applicator maps these actions to updated segmentation predictions.

- **Hybrid learning with hierarchy-aware rewards:** We combine supervised Dice and binary cross-entropy losses with MADDPG policy-gradient optimization. A region-specific reward measures segmentation improvement while penalizing hierarchical inconsistencies and unnecessary refinement steps, incorporating both pixel-level supervision and sequential decision-making into the training objective.
- **Evaluation of multi-agent segmentation refinement:** We evaluate CoSeg-MARL on 2D multimodal MRI slices from a patient-level subset of BraTS 2020, examining both zero-initialized and pretrained U-Net initialization strategies. We compare the resulting WT, TC, and ET predictions with the pretrained backbone using segmentation overlap and boundary-based metrics.

2 Related Work

Brain tumor segmentation from MRI has been widely studied using CNNs, transformers, and hybrid models to capture spatial and contextual features of heterogeneous tumor regions. Deep reinforcement learning (DRL) has also been explored for reward-guided sequential decision-making in medical image analysis, while multi-agent reinforcement learning (MARL) enables coordinated decisions among multiple agents. This section reviews these approaches, including communicative MARL, and their relationship to CoSeg-MARL.

2.1 Deep Learning for Brain Tumor Segmentation

CNNs have been central to medical image segmentation. U-Net (Ronneberger et al., 2015) introduced an encoder-decoder architecture with skip connections, while V-Net (Milletari et al., 2016) extended this approach to 3D segmentation with a Dice-based objective. DeepLabV3+ (Chen et al., 2018b) uses atrous spatial pyramid pooling for multi-scale context, and nnU-Net (Isensee et al., 2021) automatically adapts preprocessing, architecture, training, and post-processing to each task. Transformer-based models such as TransBTS (Wang et al., 2021) combine CNN features with long-range contextual modeling for multimodal brain tumor segmentation. Other approaches incorporate anatomical relationships among tumor sub-regions. Zhang et al. (2025a) proposed multi-label segmentation following the clinical hierarchy from whole tumor (WT) to tumor core (TC) and enhancing tumor (ET), with a hierarchical consistency transformation. Cascaded CNNs similarly segment these nested regions sequentially, using larger-region predictions to constrain smaller ones (Wang et al., 2019). Unlike these supervised approaches, CoSeg-MARL explores anatomical hierarchy through cooperative sequential refinement, where specialized agents exchange information and iteratively update region-specific predictions.

2.2 Deep Reinforcement Learning for Segmentation

Reinforcement learning (RL) provides a framework for formulating image segmentation as a sequential decision-making problem, in which an agent learns to refine its predictions through actions guided by reward signals. Early approaches explored classical Q-learning (Watkins & Dayan, 1992) for image segmentation (Sahba et al., 2008), but tabular methods face scalability challenges in high-dimensional state and action spaces. Deep Q-Networks (DQN) (Mnih et al., 2015) address aspects of this limitation by approximating action-value functions using deep neural networks. Related approaches include DeepOutline (Wang et al., 2018), which investigates sequential object boundary delineation, and SeedNet (Song et al., 2018), which learns to generate interaction seeds for segmentation. These methods illustrate the potential of learning sequential segmentation decisions, although discrete action representations can become restrictive when the refinement process requires high-dimensional continuous adjustments. Actor-critic reinforcement learning methods provide alternative approaches to sequential decision-making. Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2015) learns deterministic policies for continuous action spaces, while Proximal Policy Optimization (PPO) (Schulman et al., 2017) and Asynchronous Advantage Actor-Critic (A3C) (Mnih et al., 2016) provide different policy-optimization strategies. Reinforcement learning has also been investigated in medical image analysis for tasks such as anatomical landmark detection (Alansary et al., 2019) and object localization (Caicedo & Lazebnik, 2015). These studies demonstrate the applicability of sequential decision-making to image analysis tasks. However, refining

multiple anatomically related tumor regions introduces an additional coordination problem: decisions affecting one region may need to account for the predictions of other regions. This motivates investigating multi-agent reinforcement learning, in which specialized agents can coordinate their refinement decisions.

2.3 Multi-Agent Reinforcement Learning for Medical Image Analysis

Multi-agent reinforcement learning (MARL) has been investigated in medical image analysis to enable multiple agents to coordinate their decisions when operating on related anatomical structures. Vlontzos et al. (2019) introduced a multi-agent approach for anatomical landmark detection in which agents implicitly exchange information through shared convolutional representations. Extending this direction, Leroy et al. (2020) investigated communicative MARL for anatomical landmark detection in 3D medical images, combining implicit communication through shared network representations with explicit information exchange between agents. More recently, Zhang et al. (2025b) proposed a multi-agent approach for medical image segmentation based on active contour optimization, in which agents associated with contour points coordinate their movements to refine object boundaries. These studies demonstrate different forms of multi-agent coordination in medical image analysis, ranging from landmark localization to contour-based segmentation. However, they investigate different decision formulations from the joint refinement of multiple anatomically nested tumor regions, where region-specific agents must coordinate their predictions while maintaining hierarchical consistency.

2.4 Positioning of CoSeg-MARL

While existing communicative MARL approaches have investigated anatomical landmark detection (Vlontzos et al., 2019; Leroy et al., 2020) and coordinated contour optimization (Zhang et al., 2025b), CoSeg-MARL investigates a different formulation in which three agents specialize in the anatomically nested WT, TC, and ET regions and cooperatively refine dense segmentation predictions. Rather than optimizing landmark locations or individual contour points, each agent generates continuous global controls and spatial refinement coefficients that are mapped to updated segmentation predictions through a deterministic applicator. The framework combines shared observations, shared convolutional representations, and explicit message passing with a hybrid supervised and MADDPG-based training objective. Its reward incorporates segmentation improvement, hierarchical inconsistency penalties, and a per-step refinement cost. Together, these components formulate brain tumor segmentation as a cooperative sequential refinement problem that integrates region-specific decisions with anatomical hierarchy.

3 Methodology

This section presents CoSeg-MARL, a communicative multi-agent reinforcement learning framework for brain tumor segmentation from multimodal MRI. We formulate segmentation as a cooperative sequential decision-making problem and employ Multi-Agent Deep Deterministic Policy Gradient (MADDPG) to train three specialized agents, each responsible for refining one of the anatomically nested tumor regions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The agents coordinate through shared observations, shared convolutional representations, and explicit message passing. The proposed pipeline consists of data preprocessing, a pretrained segmentation backbone, the multi-agent reinforcement learning formulation, the communicative actor-critic architecture, and the learning and termination procedures. An overview of the framework is illustrated in Figure 1.

3.1 Data Preparation

We use the BraTS 2020 dataset (Menze et al., 2014), comprising 369 preoperative glioma cases, including 293 high-grade gliomas (HGG) and 76 low-grade gliomas (LGG). Each case contains four co-registered MRI modalities: native T1-weighted (T1), contrast-enhanced T1-weighted (T1Gd), T2-weighted (T2), and fluid-attenuated inversion recovery (FLAIR), with image dimensions of $240 \times 240 \times 155$ voxels. The dataset provides expert-annotated segmentation masks identify-

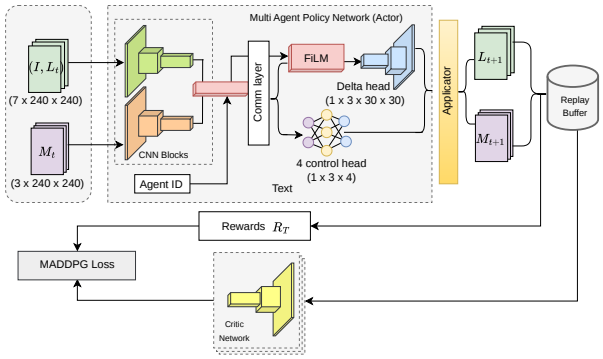


Figure 1: Overview of CoSeg-MARL, where three agents coordinate through shared representations and message passing to iteratively refine brain tumor segmentation using continuous actions, a deterministic applicator, and hybrid supervised–MADDPG learning

ing enhancing tumor (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor core (NCR/NET).

From the full dataset, we select a subset of 200 patients based on the number of informative 2D MRI slices available per patient. The selected patients are shuffled and partitioned at the patient level into training (65%), testing (15%), and validation (20%) sets, corresponding to 130, 30, and 40 patients, respectively. Patient-level partitioning ensures that slices from the same patient do not appear in multiple data splits.The original BraTS annotations are combined to construct three anatomically nested segmentation targets: whole tumor (WT), comprising NCR/NET, ED, and ET; tumor core (TC), comprising NCR/NET and ET; and enhancing tumor (ET). These regions satisfy the hierarchical relationship $\mathrm{ET} \subseteq \mathrm{TC} \subseteq \mathrm{WT}$. The processed data set undergoes per-modality z-score normalization over brain voxels only. This makes it scale invariant. The normalization is performed as: $\mathrm{I}^i_{norm} = \frac{I^i_{m>0} - \mu_i}{\sigma_i}, where I^i_{m>0}$ is the raw image for each modality i with background masked, μ_i is the mean and σ_i is the standard deviation of the foreground image. The multi modal MRI input preserves key structural information that supports the learning of spatial features and tissue contrasts necessary for differentiating tumor sub-regions. This normalization reduces differences in intensity scale across MRI modalities while preserving their complementary anatomical information.

3.2 Backbone Architecture

For the warm-start setting, we train a 2D U-Net on the preprocessed multimodal BraTS dataset. The network follows an encoder-decoder architecture with skip connections and provides initial segmentation predictions for subsequent refinement by the RL agents. The backbone is trained using binary cross-entropy loss and produces segmentation logits for the three target regions: WT, TC, and ET.We evaluate the framework under both warm-start and cold-start settings. In the warm-start setting, the pretrained backbone is frozen, and its output logits initialize the segmentation masks subsequently refined by the RL agents. In the cold-start setting, the agents begin with zero-initialized logits rather than pretrained segmentation predictions.

3.3 Reinforcement Learning Formulation

We formulate brain tumor segmentation as a cooperative multi-agent Markov game, in which three agents are responsible for iteratively refining the whole tumor (WT), tumor core (TC), and enhancing tumor (ET) segmentation masks, respectively. The agents operate in a shared environment and coordinate through shared representations, current segmentation masks, and explicit message passing. At each time step t, the agents observe the current segmentation state and produce continuous refinement actions that determine the next state s_{t+1}. The formulation is defined by the tuple $(\mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, T, \{R_i\}_{i=1}^N, \gamma)$, where:

- **Agents ($\mathcal{N}$):** $\mathcal{N} = \{1, 2, 3\}$ denotes the three cooperative agents, each assigned to one segmentation target: WT, TC, or ET.
- **State:** At time step t, the shared environment state is defined as $s_t = (I, \mathbf{M}_t, \mathbf{L}_t)$, where I denotes the four-channel multimodal MRI input, $\mathbf{M}_t = (M_t^{(1)}, M_t^{(2)}, M_t^{(3)})$ contains the current segmentation masks for WT, TC, and ET, and $\mathbf{L}_t = (L_t^{(1)}, L_t^{(2)}, L_t^{(3)})$ contains their corresponding logit maps. Each agent uses this shared information to refine its assigned tumor region.
- **Action:** At each time step t, agent i produces a 904-dimensional continuous action $a_t^{(i)} \in \mathcal{A}_i$, comprising four global controls and a 30×30 spatial refinement grid. All action components are bounded to $[-1, 1]$ using the $\tanh$ activation function. The 4 global controls are (1) Threshold_shift (g1): Adjusts the foreground–background decision threshold. (2) Logit_temperature (g2): Controls the scaling of the segmentation logits. (3) Morphology (g_3): Controls morphological expansion or contraction of the segmentation mask. (4) Hierarchy_strength (g4): Controls enforcement of the anatomical hierarchy $ET \subseteq TC \subseteq WT$. This coarse grid represents local adjustments to the segmentation prediction and is upsampled to the original 240×240 image resolution by the deterministic action applicator.
- **Transition and discount factor:** T denotes the environment transition induced by the joint action $\mathbf{a}_t = (a_t^{(1)}, a_t^{(2)}, a_t^{(3)})$, which updates the shared segmentation state from s_t to s_{t+1}. The discount factor $\gamma \in [0, 1]$ controls the relative importance of future rewards.
- **Reward:** Each agent receives a region-specific reward based on the improvement in Dice coefficient for its assigned segmentation target, together with an anatomical hierarchy violation penalty and a small per-step cost. The reward function R_i assigns the following clipped reward $r_t^{(i)}$ to agent i at time step t:

$$r_t^{(i)} = \text{clip}\left(\text{Dice}(M_{t+1}^{(i)}, G^{(i)}) - \text{Dice}(M_t^{(i)}, G^{(i)}) - 0.5\frac{v_t}{1000} - 0.01,\, -1,\, 1\right). \quad (1)$$

  Here, $G^{(i)}$ denotes the ground-truth mask for agent i, and v_t denotes the number of anatomical hierarchy violations at time step t.

3.4 CoSeg-MARL Architecture and Agent Coordination

The proposed framework is implemented using the Multi-Agent Deep Deterministic Policy Gradient (MADDPG) algorithm (Lowe et al., 2017), which extends deterministic actor-critic learning to multi-agent settings through centralized training and decentralized execution. Each agent is assigned to refine one of the three segmentation targets: WT, TC, or ET.

3.4.1 Actor Network

The three agents share a common actor network with a shared encoder and are distinguished through learned agent embeddings, agent-specific communication parameters, and feature-wise linear modulation (FiLM). The actor follows a U-Net-style encoder-bottleneck architecture, replacing the conventional full-resolution segmentation decoder with a lightweight decoder and two action heads. The actor processes a seven-channel input comprising the four MRI modalities and three current segmentation logit maps, together with the three current segmentation masks through a separate convolutional pathway. The encoded representation is concatenated and compressed into a 256-channel bottleneck. A learned agent embedding conditions the encoded representation. The communication layer aggregates messages across the three agents and concatenates the pooled message with each agent's features, allowing decisions to incorporate information from the other agents. Agent-conditioned FiLM parameters scale and shift the shared bottleneck features. The resulting representation is passed to two action heads: a fully connected head producing four global controls and a decoder head producing a 30×30 spatial refinement grid. Both outputs are bounded to $[-1, 1]$ using $\tanh$ activations. The actor is optimized using a hybrid objective combining supervised Dice loss, binary cross-entropy (BCE), and a reinforcement learning term derived from the critic.

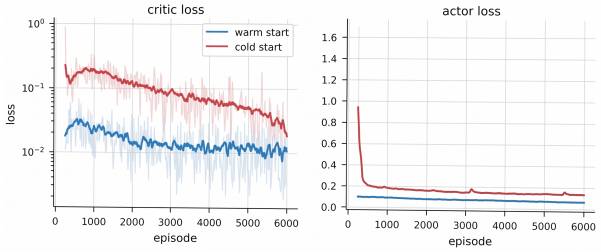


Figure 2: Actor and critic losses over 6,000 episodes for warm- and cold-start refinement.

3.4.2 Critic Network

Each agent has a separate critic that estimates the Q-value of its refinement action using the shared segmentation state and the agent's own action. The critic processes the four-channel MRI input and the three current logit maps through convolutional layers, producing a 30×30 feature representation. This representation is concatenated with the agent's 30×30 spatial refinement action and processed by additional convolutional layers. The resulting features are combined with the four global action controls and passed through fully connected layers to produce a scalar Q-value.

3.4.3 Cooperative Mechanism

The segmentation targets follow the anatomical hierarchy ET $\subseteq$ TC $\subseteq$ WT. To reduce anatomically inconsistent predictions, CoSeg-MARL uses three coordination mechanisms: **Shared convolutional representations** enable agents to learn common MRI features with fewer parameters; **shared observations** provide each agent with all agents' current masks and logits; and **explicit message passing** aggregates messages from all three agents and concatenates the pooled message with each agent's features before the action heads.

3.5 Learning Process and Termination

Training is conducted under two refinement settings. In the warm-start setting, the agents begin with segmentation logits generated by the frozen pretrained backbone. In the cold-start setting, refinement begins with zero-initialized logit maps. In both settings, the agents iteratively update the segmentation predictions, and the resulting rewards are used to optimize the actor and critic networks. Training continues until one of two termination conditions is met: (1) the Dice score exceeds a high-precision threshold, or (2) a maximum number of decision steps is reached. This iterative mechanism emulates how expert radiologists adjust segmentation over multiple assessments. The experiments are conducted on a workstation equipped with a NVIDIA Tesla V100 GPU with 32 GB memory. The model is trained for 6000 episodes using the Adam optimizer with a mini-batch size of 32. The actor learning rate is $3 \cdot 10^{-4}$ while the critic learning rate is $2 \cdot 10^{-4}$. The discount factor γ is set to 0.95, with τ of 0.005 and action noise of 0.2 for exploration. Although our implementation primarily utilizes MADDPG, the framework is compatible with other off-policy RL algorithms such as Soft Actor-Critic (SAC) Haarnoja et al. (2018) and Twin Delayed DDPG (TD3) Fujimoto et al. (2018), depending on the desired trade-offs in stability and exploration.

4 Experimental Evaluation

We evaluate CoSeg-MARL on the BraTS 2020 dataset through qualitative and quantitative analyses. Predicted segmentation masks are compared with expert-annotated ground truth using the Dice Similarity Coefficient (DSC), Hausdorff Distance (HD), average symmetric surface distance (ASSD), and relative volume difference (RVD). All metrics are computed on individual 2D slices and averaged across slices, rather than aggregated into patient-level volumetric scores. These metrics characterize segmentation overlap, boundary agreement, and differences in segmented region size.

Figure 2 compares actor and critic training losses for warm- and cold-start refinement over 6,000 episodes. The critic loss is shown on a logarithmic scale and the actor loss on a linear scale. The different loss trajectories illustrate the effect of initialization on training.

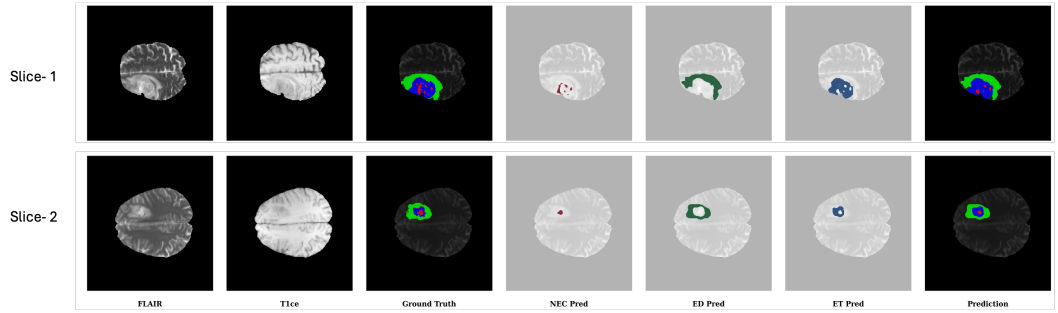


Figure 3: Qualitative segmentation on representative 2D MRI slices, showing MRI inputs, ground truth, predicted tumor components, and prediction overlays.

Figure 3 presents qualitative segmentation results on representative MRI slices. The predicted tumor components are shown alongside the ground-truth annotations and MRI inputs, illustrating their spatial agreement and the delineation of different tumor regions.

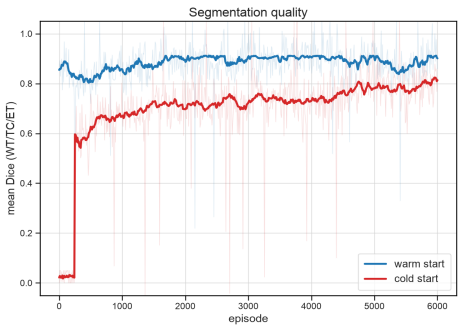


Figure 4: Mean Dice coefficient across WT, TC, and ET during training over 6,000 episodes for warm-start and cold-start refinement.

Figure 4 shows the mean Dice progression across WT, TC, and ET during 6,000 training episodes. Cold-start refinement begins near zero and improves during training, whereas warm-start refinement begins with higher Dice scores and improves more gradually. The reported mean Dice scores are 0.821 and 0.916, respectively. For quantitative evaluation, we report segmentation overlap, boundary agreement, and region-size differences using the following metrics:

- **Dice Similarity Coefficient (DSC)** measures the spatial overlap between predicted and ground truth masks:

$$\mathrm{DSC} = \frac{2|S_{\mathrm{gt}} \cap S_{\mathrm{pred}}|}{|S_{\mathrm{gt}}| + |S_{\mathrm{pred}}|}$$

  A higher DSC indicates greater overlap between the predicted and ground-truth regions. DSC is computed using binarized segmentation masks throughout the evaluation..

- **95th-percentile Hausdorff Distance (HD95)** measures boundary disagreement using the 95th percentile of the directed distances between predicted and ground-truth contours:

$$HD_{95}(S_{\mathrm{gt}}, S_{\mathrm{pred}}) = \max\left\{d_{95}(S_{\mathrm{pred}}, S_{\mathrm{gt}}), d_{95}(S_{\mathrm{gt}}, S_{\mathrm{pred}})\right\}.$$

  Lower HD values imply more precise boundary localization.

- **Relative Volume Difference (RVD)** measures the relative difference in segmented region size between the prediction and ground truth. In our 2D evaluation, region size is measured

Table 1: Comparison of backbone, cold-start, and warm-start refinement.

Method	Metric	WT	TC	ET	Mean
Backbone	Dice	0.811 ± 0.083	0.830 ± 0.028	0.838 ± 0.025	0.826
	IoU	0.749 ± 0.110	0.782 ± 0.300	0.778 ± 0.260	0.769
	BoundF	0.578 ± 0.248	0.420 ± 0.311	0.643 ± 0.169	0.547
	HD	14.98 ± 12.48	4.90 ± 5.07	4.74 ± 2.07	8.20
	ASSD	6.62 ± 4.14	2.74 ± 2.37	3.66 ± 3.78	4.34
	RVD	0.455 ± 3.76	0.649 ± 4.12	0.394 ± 2.38	0.499
Cold	Dice	0.873 ± 0.039	0.827 ± 0.110	0.763 ± 0.065	0.821
	IoU	0.551 ± 0.327	0.730 ± 0.197	0.645 ± 0.315	0.642
	BoundF	0.464 ± 0.315	0.673 ± 0.311	0.743 ± 0.169	0.627
	HD	14.62 ± 13.60	4.80 ± 4.07	5.33 ± 3.07	8.25
	ASSD	4.62 ± 4.23	1.74 ± 1.58	1.66 ± 1.61	2.67
	RVD	0.774 ± 2.96	0.382 ± 3.840	0.228 ± 1.893	0.46
Warm	Dice	0.927 ± 0.036	0.916 ± 0.039	0.907 ± 0.069	0.916
	IoU	0.725 ± 0.210	0.789 ± 0.127	0.776 ± 0.160	0.764
	BoundF	0.732 ± 0.201	0.826 ± 0.173	0.879 ± 0.006	0.812
	HD	13.94 ± 11.16	4.44 ± 3.61	3.13 ± 1.39	7.17
	ASSD	3.03 ± 2.70	1.20 ± 1.08	0.79 ± 0.36	1.67
	RVD	0.228 ± 1.894	0.331 ± 5.202	0.642 ± 6.647	0.40

Table 2: Quantitative comparison of segmentation methods on BraTS.

Method	DSC	SD	HD	SD	RVD	ABD
Grab-Cut (Li et al., 2018)	78.41	15.62	21.52	11.27	4.12	2.56
PSPNet (Yan et al., 2021)	75.49	9.41	24.58	15.26	4.71	2.89
FCN (Kumar et al., 2018)	82.37	5.56	19.64	19.79	6.06	2.39
U-Net (Ronneberger et al., 2015)	84.71	6.52	15.92	6.85	2.40	1.89
V-Net (Milletari et al., 2016)	85.29	6.82	16.78	6.60	3.49	2.02
DeepLabV3+ (Chen et al., 2018a)	86.45	5.09	23.08	19.07	-6.18	2.20
RL-Based (Sahba et al., 2008)	93.69	1.04	14.00	6.37	-0.30	1.80
MARL-MambaContour (Zhang et al., 2025b)	91.74	–	–	–	–	–
CoSeg-MARL	**91.60**	**4.80**	**7.17**	**5.39**	**0.40**	**1.67**

in pixels:

$$\text{RVD} = 100 \times \left(\frac{|S_{pred}|}{|S_{gt}|} - 1 \right)$$

- **Average Symmetric Surface Distance (ASSD)** measures the mean bidirectional distance between the predicted and ground-truth boundaries.

$$ASSD(S_{pred}, S_{gt}) = \frac{\sum_{p \in \partial S_{pred}} d(p, \partial S_{gt}) \; + \; \sum_{q \in \partial S_{gt}} d(q, \partial S_{pred})}{|\partial S_{pred}| + |\partial S_{gt}|}$$

Table 1 illustrates the ablation study of the proposed methodology on BraTS dataset using two different refinement strategy. (1) Refinement from scratch - cold start, logits and masks intialized with all zeros and RL refines over this at each step shows mean DSC of 0.821. (2) Refinement from pre-computed logits from backbone - warm start, where RL model refines over the intial UNet logits and achieves DSC of 0.916.

Table 2, presents a comprehensive quantitative comparison of segmentation performance across various established methods on the BraTS dataset. It reports segmentation accuracy across various metrics on held-out patients. The proposed CoSeg-MARL achieves 91.60% mean DSC with 2.64% stdandard deviation, 7.17mm for mean HD with 5.39mm standard deviation, 0.42 RVD, 1.67 ASSD compared with ABD, 0.812 mean boundary F-score. These results underscore the advantage of hierarchical, cooperative multi-agent reinforcement learning in modeling complex tumor sub-regions, providing improved spatial accuracy and consistency compared to traditional segmentation architectures.

5 Conclusion

We introduced CoSeg-MARL, a communicative multi-agent reinforcement learning framework for iterative brain tumor segmentation. Three specialized agents coordinate through shared representations and message passing to refine WT, TC, and ET predictions using hybrid supervised and reinforcement learning. On BraTS 2020, warm-start refinement achieved a mean Dice score of 0.916, compared with 0.821 for cold-start refinement, and improved Dice and surface-distance metrics relative to the pretrained U-Net backbone. While these results support the complete warm-start framework, the individual effects of communication, reinforcement learning, and hierarchical constraints remain to be isolated. Future work will investigate these components through controlled ablations and evaluate the framework on additional datasets and volumetric segmentation tasks.

AI USE STATEMENT

Generative AI tools were used to improve writing, identify relevant literature, explore methodological concepts, and assist with code completion. They were not used to conceive the research idea, design the methodology, conduct experiments, interpret results, formulate claims, or provide substantive authorship. All AI-assisted material was reviewed and verified by the authors, who take full responsibility for the paper's content, results, and claims.

REPRODUCIBILITY STATEMENT

To support reproducibility, we detail the data preprocessing, backbone architecture, reinforcement learning formulation, network architectures, training procedure, and termination criteria. The Evaluation section reports the experimental settings, evaluation metrics, and ablation studies.

REFERENCES

Amir Alansary, Ozan Oktay, Wenjia Li, et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Medical Image Analysis*, 2019.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):170117, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 2018a.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *European Conference on Computer Vision (ECCV)*, 2018b.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. *arXiv preprint arXiv:1802.09477*, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, feb 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z.

Sanjay Kumar, Ashish Negi, JN Singh, and Himanshu Verma. A deep learning for brain tumor mri images semantic segmentation using fcn. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–4. IEEE, 2018.

Guy Leroy, Daniel Rueckert, and Amir Alansary. Communicative reinforcement learning agents for landmark detection in brain images, 2020. URL `https://arxiv.org/abs/2008.08055`.

Yubing Li, Jinbo Zhang, Peng Gao, Liangcheng Jiang, and Ming Chen. Grab cut image segmentation based on image region. In *2018 IEEE 3rd international conference on image, vision and computing (ICIVC)*, pp. 311–315. IEEE, 2018.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, et al. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, Cynthia Hawkins, HK Ng, Stefan M Pfister, Guido Reifenberger, et al. The 2021 who classification of tumors of the central nervous system: a summary. *Neuro-oncology*, 23(8):1231–1251, 2021.

Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pp. 6379–6390, 2017.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 2015.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, et al. Asynchronous methods for deep reinforcement learning. *International Conference on Machine Learning (ICML)*, 2016.

Novsheena Rasool and Javaid Iqbal Bhat. A critical review on segmentation of glioma brain tumor and prediction of overall survival. *Archives of Computational Methods in Engineering*, 32(3): 1525–1569, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015.

Farshid Sahba, Anastasios N Venetsanopoulos, and Douglas Bird. Reinforcement learning-based medical image segmentation. *Artificial Intelligence in Medicine*, 2008.

John Schulman, Filip Wolski, Prafulla Dhariwal, et al. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Haoyan Song, Xiang Bai, et al. Seednet: Automatic seed generation with deep reinforcement learning for interactive segmentation. *IEEE Transactions on Image Processing*, 2018.

Athanasios Vlontzos, Amir Alansary, Konstantinos Kamnitsas, Daniel Rueckert, and Bernhard Kainz. Multiple landmark detection using multi-agent reinforcement learning, 2019. URL `https://arxiv.org/abs/1907.00318`.

Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Frontiers in Computational Neuroscience*, 13:56, 2019. doi: 10.3389/fncom.2019.00056.

Jin Wang, Wei Zhang, Ying Xu, et al. Deepoutline: An end-to-end deep reinforcement learning framework for object outlining. *arXiv preprint arXiv:1803.01215*, 2018.

Wenxuan Wang, Chen Chen, Meng Ding, Hong Yu, Sen Zha, and Jiangyun Li. Transbts: Multimodal brain tumor segmentation using transformer. In *International conference on medical image computing and computer-assisted intervention*, pp. 109–119. Springer, 2021.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 1992.

Lingfei Yan, Dawei Liu, Qi Xiang, Yang Luo, Tao Wang, Dali Wu, Haiping Chen, Yu Zhang, and Qing Li. Psp net-based automatic segmentation network model for prostate magnetic resonance imaging. *Computer Methods and Programs in Biomedicine*, 207:106211, 2021.

Pengfei Zhang, Jinke Wang, Yuanzhi Cheng, and Shinichi Tamura. Clinical learning order-guided deep neural network for brain tumor segmentation. *IEICE Transactions on Information and Systems*, E108.D(6):615–628, 2025a. doi: 10.1587/transinf.2024EDP7222.

Ruicheng Zhang, Yu Sun, Zeyu Zhang, Jinai Li, Xiaofan Liu, Au Hoi Fan, Haowei Guo, and Puxin Yan. Marl-mambacontour: Unleashing multi-agent deep reinforcement learning for active contour optimization in medical image segmentation, 2025b. URL https://arxiv.org/abs/2506.18679.