

ARTICLE

Oracle-Conditioned Deep Q-Network Refinement of Small Brain Tumor Components in Axial Magnetic Resonance Imaging

Bakhytzhan Omarov¹, Balnur Kenjayevea^{2,*}, Daniyar Sultan^{3,*} and Zhanseri Ikram³

¹School of Physical Education and Sports, International University of Tourism and Hospitality, Turkistan, Kazakhstan

²School of Languages, International University of Tourism and Hospitality, Turkistan, Kazakhstan

³School of Digital Technologies, Narxoz University, Almaty, Kazakhstan

*Corresponding Authors: Balnur Kenjayevea. Email: balnur.kendjaeva@iuth.edu.kz;

Daniyar Sultan. Email: daniyar.sultan@narxoz.kz

Received: 09 July 2026; Accepted: 10 September 2026; Published: 28 September 2026

ABSTRACT: Brain tumor localization in magnetic resonance imaging remains challenging for small, low-contrast lesions. This study presents Small Brain Lesion Deep Q Network (SBL-DQN), a two-dimensional lesion-wise reinforcement-learning framework that refines an axis-aligned region of interest on a co-registered axial multi-channel MRI slice. In the current evaluation protocol, voxel reference masks are resampled, processed by 26-connected-component analysis, and filtered using a maximum in-plane diameter of 10 mm; one axial episode is then executed for each reference-enumerated component. The mask and component coordinates are not included in the DQN state, but ground-truth annotations are required to generate the evaluation candidate list and to compute reward traces and localization metrics. Accordingly, the reported experiment is oracle-conditioned lesion-wise localization refinement rather than autonomous lesion detection. The agent uses seven actions (up, down, left, right, zoom in, zoom out, and stop) and does not move through the anatomical z-axis. On the internal BraTS 2025 evaluation, the complete configuration yielded 97.46% localization accuracy, 97.23% precision, 96.88% recall, 97.05% F1-score, mean Intersection over Union of 0.834, AP@0.5 of 0.961, and center localization error of 0.91 mm. BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE remain qualitative transfer examples. Component-wise ablation, eligible-lesion accounting, and a balanced failure-case archive were not retained in the available experimental record; therefore, no individual component contribution, autonomous detection performance, or comprehensive robustness claim is made.

KEYWORDS: Oracle-conditioned localization; lesion-wise ROI refinement; deep reinforcement learning; deep Q network (DQN); small brain tumor components; magnetic resonance imaging; medical image analysis

1 Introduction

Brain tumors remain one of the most challenging neurological disorders because their early detection and accurate localization are essential for diagnosis, treatment planning, surgical navigation, and post-treatment monitoring [1]. Magnetic resonance imaging (MRI) is the primary imaging modality for brain tumor assessment owing to its excellent soft-tissue contrast and ability to visualize anatomical structures using multiple imaging sequences [2]. Nevertheless, the localization of small brain tumor lesions remains particularly difficult due to their limited size, irregular morphology, heterogeneous intensity distribution, and similarity to surrounding healthy tissues. These characteristics often reduce the sensitivity of conventional computer-aided diagnosis systems and increase the likelihood of missed detections, especially during the early stages of disease progression [3]. Consequently, the development of intelligent localization

approaches capable of accurately identifying small lesions has become an important research direction in medical image analysis and clinical decision support systems [4]. Recent advances in artificial intelligence have demonstrated considerable potential for improving the efficiency and reliability of automated brain tumor analysis, thereby supporting radiologists in routine clinical practice.

Deep learning has substantially advanced medical image interpretation by enabling automatic extraction of hierarchical image representations without the need for handcrafted features. Convolutional neural networks, encoder-decoder architectures, attention mechanisms, and transformer-based models have achieved remarkable performance in various brain tumor segmentation and classification tasks [5,6]. Despite these successes, many existing approaches rely on exhaustive dense prediction or sliding-window strategies that require significant computational resources and frequently struggle to localize very small lesions with high precision [7]. Moreover, these supervised models generally treat localization as a static prediction problem and do not explicitly learn sequential search behaviors that progressively refine lesion positions. Reinforcement learning provides an attractive alternative by formulating localization as a sequential decision-making process in which an intelligent agent continuously interacts with an imaging environment to identify the optimal lesion location through a sequence of actions [8]. Such an adaptive search strategy can substantially reduce unnecessary image exploration while improving localization efficiency and robustness.

Deep Q Networks (DQNs) integrate deep neural function approximation with Q-learning, providing a principled framework for learning sequential localization policies directly from high-dimensional medical images. The application of DQN-based strategies to brain tumor localization, however, predates the present study. Stember and Shalu [9] investigated a DQN/Q-learning formulation using 30 two-dimensional contrast-enhanced BraTS slices for training and an independent set of 30 slices for evaluation, with each image containing a single lesion. Their formulation represented localization as navigation within a 60×60 -pixel gridworld using three discrete actions, namely stay, move down, and move right, and considered localization successful when the agent reached a grid cell overlapping the lesion. In contrast, the proposed SBL-DQN formulates the task as variable-size ROI refinement over co-registered axial multi-channel MRI and employs seven actions for bidirectional translation, zooming, and termination. Its state representation combines the current ROI patch with ROI geometry, normalized episode progress, and previous-action information, while the final prediction comprises a two-dimensional bounding box and lesion centroid. Importantly, the evaluation protocol uses reference masks to enumerate lesion components and identify component-intersecting axial slices before individual episodes are initialized. Accordingly, SBL-DQN is evaluated as an oracle-conditioned lesion-wise localization and refinement framework rather than an autonomous lesion detection system. Its contribution therefore lies in the proposed sequential ROI refinement formulation and its controlled internal evaluation.

2 Related Works

Automatic analysis of brain tumors from magnetic resonance imaging has received significant attention due to its potential to improve diagnostic accuracy and reduce the workload of radiologists. Early computer-aided diagnosis systems primarily relied on handcrafted features describing texture, intensity, edge information, and shape characteristics, followed by conventional machine learning classifiers. Although these approaches achieved reasonable performance under controlled conditions, their ability to generalize across different imaging protocols and tumor characteristics remained limited [10]. The emergence of deep learning has fundamentally transformed brain tumor analysis by enabling end-to-end feature learning directly from medical images [11]. Convolutional neural networks have become the dominant architecture for tumor detection, classification, and segmentation because of their strong capability to extract hierarchical spatial representations from MRI data [12]. Multi-scale feature extraction strategies

have further enhanced the detection of lesions exhibiting substantial variations in size and appearance [13]. More recently, transformer-based architectures have demonstrated promising performance by modeling long-range contextual dependencies that complement conventional convolutional operations.

Brain tumor localization has traditionally been addressed using object detection and semantic segmentation frameworks. Region proposal networks, two-stage detectors, and single-stage detection algorithms have been successfully adapted to identify tumor regions within MRI scans [14]. Encoder-decoder segmentation networks have also been widely employed to generate dense lesion masks while simultaneously providing localization information [15]. Several studies have incorporated attention mechanisms to emphasize diagnostically relevant regions and suppress background structures, thereby improving localization precision [16–18]. Hybrid convolution-transformer architectures have further improved feature representation by combining local spatial information with global contextual modeling [19]. Despite these advances, accurate localization of small brain tumor lesions remains challenging because of low lesion contrast, irregular boundaries, substantial anatomical variability, and severe class imbalance between tumor and healthy tissues [20]. Consequently, existing supervised learning approaches often experience reduced sensitivity when detecting small lesions occupying only a limited portion of the brain volume.

Reinforcement learning has been used for anatomical landmark localization, lesion search, and medical-image navigation [8,21,22]. Direct precedent exists for brain tumor localization: Stember and Shalu [9] showed that a DQN/Q-learning agent could localize one lesion on a two-dimensional contrast-enhanced MRI slice using a coarse gridworld and a very small training set. That study output a lesion-overlapping grid location, whereas SBL-DQN refines a variable-size 2D bounding box by translation and zoom actions. SBL-DQN further includes ROI geometry and action history in the state and combines overlap, centroid-distance, size, step, and terminal terms in the reward. These are formulation-level distinctions, not evidence that any individual element independently causes the reported improvement.

Although dense segmentation and detection networks provide comprehensive spatial predictions, they process large portions of the image and optimize a different output from a sequential localization policy [6,13,23,24]. The present study therefore evaluates SBL-DQN primarily against reinforcement learning baselines on the same lesion-wise localization task. The intended contribution is a reproducible 2D ROI-refinement formulation for small-lesion localization, with explicit limits on dimensionality, multi-lesion inference, and external validation.

3 Materials and Methods

This study presents SBL-DQN (Small Brain Lesion Deep Q Network), a two-dimensional reinforcement-learning framework for oracle-conditioned lesion-wise localization on axial MRI slices. The input is a co-registered multi-channel axial slice, represented as $C \times H \times W$, with T1, T1ce, T2, and FLAIR channels when available. Each episode operates within one axial plane; neither the state transition nor the action space changes the anatomical slice index. MRI volumes organize subjects and supply reference lesion components, whereas the DQN observation and predicted bounding box are two-dimensional. At evaluation, the candidate list is generated from the ground-truth mask rather than from an image-only proposal model; the method must therefore not be interpreted as autonomous lesion detection.

Fig. 1 separates candidate construction from ROI refinement. For training and internal evaluation, each resampled voxel reference mask is converted to the whole-tumor union, decomposed using three-dimensional 26-connectivity, and filtered by the prespecified maximum in-plane diameter criterion of 10 mm. An axial slice intersecting each eligible component is selected, and one brain-field-initialized episode is created for that assigned component. No learned proposal network, thresholded image-only detector, or unsupervised connected-component generator is used at test time. The DQN state excludes the mask and

component coordinates, but the evaluation harness requires the reference annotation to enumerate targets and compute the reward trace and outcome metrics. Consequently, all quantitative results in this manuscript describe oracle-conditioned lesion-wise refinement, not autonomous examination-level detection.

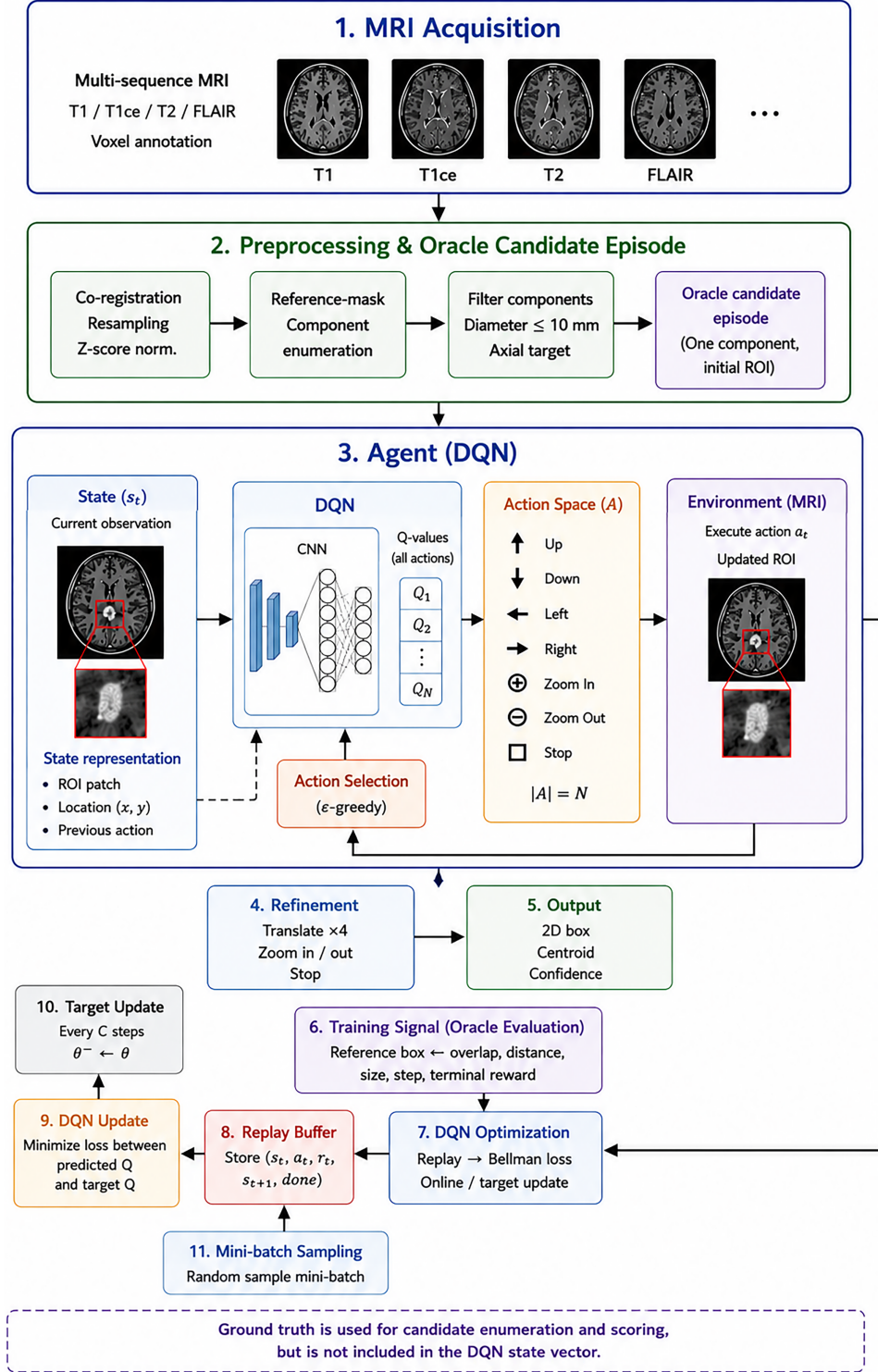


Figure 1: Oracle-conditioned workflow of SBL-DQN for reference-defined axial brain tumor component localization.

Prior to reinforcement learning, all MRI volumes underwent identical preprocessing procedures to reduce inter-subject variability. Skull stripping was first performed to eliminate non-brain tissues, followed by N4 bias field correction to compensate for low-frequency intensity inhomogeneity. Each MRI volume was subsequently registered to a common anatomical space and resampled to an isotropic voxel spacing of $1 \text{ mm} \times 1 \text{ mm} \times 1 \text{ mm}$. Intensity normalization was then applied using Z-score normalization

$$I_{norm} = \frac{I - \mu}{\sigma}, \quad (1)$$

where I denotes the original voxel intensity, while μ and σ represent the mean and standard deviation of brain voxels, respectively. Data augmentation was further applied during training, including random rotations, horizontal and vertical flipping, scaling, Gaussian noise injection, elastic deformation, and intensity perturbation to improve model robustness.

Unlike conventional supervised localization methods, SBL-DQN formulates lesion localization as a Markov Decision Process (MDP)

$$M = (S, A, P, R, \gamma), \quad (2)$$

where S denotes the state space, A represents the action space, P is the state transition probability, R is the reward function, and γ is the discount factor. At each time step, the reinforcement learning agent observes the current search region and selects an action that maximizes the expected cumulative reward.

The state representation preserves local appearance, two-dimensional ROI geometry, normalized episode progress, and action history. Each state is defined as

$$s_t = \{I(\text{ROI}), x_t, y_t, w_t, h_t, \tau_t, a_{t-1}\}, \quad (3)$$

where $I(\text{ROI}) \in \mathbb{R}^{C \times H \times W}$ is the current multi-channel axial ROI patch; (x_t, y_t) is the in-plane ROI center; w_t and h_t are its width and height; $\tau_t = t/T$ is normalized episode progress; and a_{t-1} is the previous action. No z-coordinate or ROI depth is included because the environment is two-dimensional.

The action space contains seven discrete actions: Move Up, Move Down, Move Left, Move Right, Zoom In, Zoom Out, and Stop. Move Forward and Move Backward are not implemented.

Translation actions update the in-plane ROI center, and zoom actions change its width and height while preserving the axial slice. Coordinates are clipped to the brain-image boundary. Stop terminates the episode and returns the final 2D bounding box (x, y, w, h) , centroid (x, y) , and confidence score P^* .

Fig. 2 details the interaction loop for one reference-enumerated target. The episode begins on the target-intersecting axial slice with a standardized ROI covering the brain field. The agent observes the multi-channel patch, 2D geometry, normalized step index, and previous-action code; it then estimates seven Q-values and executes an in-plane translation, zoom, or stop action. During evaluation, network weights are frozen and actions are selected greedily. The reference target is not an observed state feature, but the evaluation harness uses it to calculate the reported reward trace and success metrics. The episode ends when Stop is selected or 300 actions have been executed.

The reward function is one of the key components of the proposed SBL-DQN framework because it directly guides the search behavior of the reinforcement learning agent. Rather than relying on a single localization criterion, a composite reward was designed to simultaneously optimize localization accuracy, convergence speed, and search efficiency. The reward at each interaction step is defined as

$$R_t = \alpha R(\text{IoU}) + \beta R(\text{Dist}) + \kappa R(\text{Size}) + \delta R(\text{Step}) + \eta R(\text{Stop}), \quad (4)$$

where α , β , κ , δ , and η are reward weights. The symbol κ replaces the earlier use of γ as a reward coefficient; γ is reserved exclusively for the MDP discount factor.

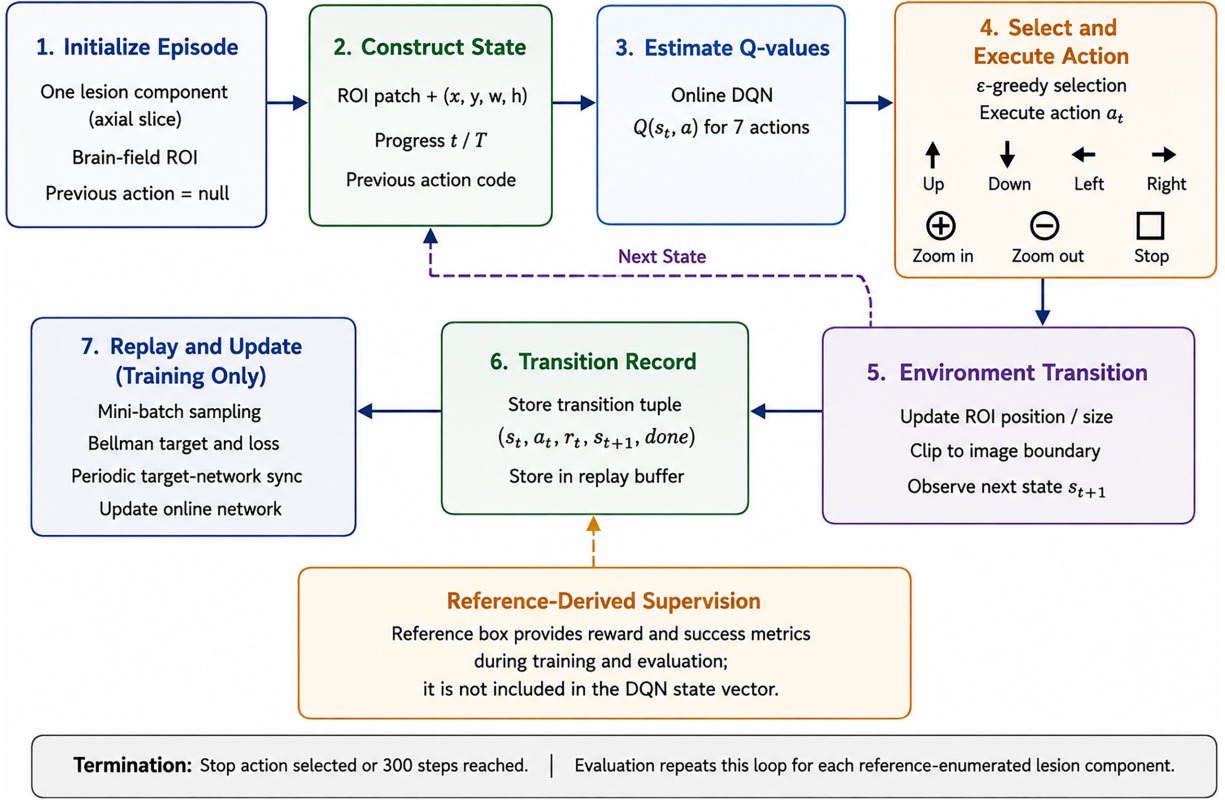


Figure 2: Oracle-conditioned lesion-wise agent–environment interaction on one axial MRI slice.

The Intersection-over-Union reward encourages progressive overlap between the predicted search window B_p and the ground-truth lesion bounding box B_g ,

$$R_{IoU} = IoU(B_p, B_g), \quad (5)$$

where,

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}, \quad (6)$$

To further improve localization precision, a centroid distance reward is introduced

$$R_{Dist} = -\frac{\|C_p - C_g\|}{D}, \quad (7)$$

where C_p and C_g are the predicted and ground-truth 2D centroids and D is the in-plane image diagonal. The size term is $R_{Size} = -\min(1, |A(B_p) - A(B_g)|/[A(B_g) + \epsilon])$, where $A(\cdot)$ is bounding-box area and ϵ prevents division by zero. Thus, the term penalizes scale mismatch without introducing a depth component.

To encourage efficient exploration, each localization step receives a small penalty

$$R_{Step} = -\lambda, \quad (8)$$

which discourages unnecessarily long trajectories. The terminal term is $R_{\text{Stop}} = +1$ when $\text{IoU} \geq 0.50$ and $\text{CLE} \leq 5 \text{ mm}$ and $R_{\text{Stop}} = -1$ otherwise. The numerical reward weights, step penalty, translation increment, zoom factor, and oracle candidate source are reported in [Table 1](#).

Table 1: Principal SBL-DQN training, environment, reward, and timing settings.

Parameter	Setting
Input	Co-registered 2D axial ROI; $C \times H \times W$ channels
State vector	ROI patch; (x, y); (w, h); normalized step $\tau = t/T$; previous action
Action space	7 actions: up, down, left, right, zoom in, zoom out, stop
Trainable parameters	4.1 million
Optimizer	Adam
Learning rate	1×10^{-4}
Mini-batch size	32 transitions
Replay-buffer capacity	50,000 transitions
Target-network update	Every 1000 environment steps
Discount factor γ	0.99
Exploration schedule	$\epsilon_{\text{max}} = 1.0$ to $\epsilon_{\text{min}} = 0.05$; exponential decay $k = 0.0025$ per episode
Training episodes	2000 episodes
Maximum episode length	300 actions
Initial ROI	Brain-field bounding rectangle on the selected axial slice
Translation increment	10% of current ROI width/height
Zoom factors	Zoom in = 0.80; zoom out = 1.25
Boundary handling	Clip ROI center and extent to image bounds
Reward weights	$\alpha = 1.0$, $\beta = 0.5$, $\kappa = 0.25$, $\delta = 0.10$, $\eta = 1.0$
Step penalty λ	0.01
Successful Stop	$\text{IoU} \geq 0.50$ and $\text{CLE} \leq 5 \text{ mm}$
Timing hardware/protocol	NVIDIA RTX 3090 24 GB; Intel Xeon Gold 6226R; PyTorch 2.1.0; CUDA 12.1; batch size 1; 10 warm-up runs; 100 internal-test volumes; search measured after preprocessing and oracle candidate enumeration
Evaluation candidate source	Reference-mask 26-connected components satisfying the $\leq 10 \text{ mm}$ criterion (oracle-conditioned; no learned candidate generator)
Episode-return reporting	Undiscounted, unscaled sum of implemented per-step rewards; averaged across evaluation episodes

Reward-scale verification. Because $R(\text{IoU})$ is in $[0, 1]$, $R(\text{Dist})$ is in $[-1, 0]$, $R(\text{Size})$ is in $[-1, 0]$, and $R(\text{Step}) = -0.01$, the weighted nonterminal reward is bounded by $-0.751 \leq R_t \leq 0.999$ for the coefficients in Table 1. When the terminal Stop term is applied once, the terminal-step range is $-1.751 \leq R_t \leq 1.999$. Therefore, for an episode of at most 300 actions, the undiscounted cumulative episode return cannot exceed 300.700 (and is no lower than -226.300 under these bounds). Average Episode Reward is calculated from the unscaled sum of the implemented per-step rewards; no multiplier, percentage conversion, or reward rescaling is applied.

Fig. 3 depicts the DQN used for the two-dimensional implementation. The network receives a $C \times H \times W$ axial ROI patch, extracts convolutional features, applies global average pooling, and concatenates the image vector with ROI geometry, normalized episode progress, and previous-action encoding. Fully connected layers produce one Q-value for each of the seven actions. The lesion mask, assigned component coordinates, and through-plane position are not network inputs. A separate target network is used only to construct the Bellman target during training.

$$Q(s, a; \theta), \quad (9)$$

where θ denotes the network parameters. The output layer produces one Q-value for each available action, and the action with the highest estimated value is selected according to

$$a_t = \arg \max_a Q(s_t, a; \theta). \quad (10)$$

To balance exploration and exploitation during training, an ϵ -greedy policy is adopted. The probability of selecting a random exploratory action is given by

$$\epsilon = \epsilon_{\min} + (\epsilon_{\max} - \epsilon_{\min})e^{-kt}, \quad (11)$$

where k controls the exploration decay rate. Initially, the agent performs extensive exploration, gradually transitioning toward deterministic exploitation as training progresses.

The proposed SBL-DQN utilizes experience replay to improve sample efficiency and reduce temporal correlations among consecutive observations. Each transition (s_t, a_t, r_t, s_{t+1}) is stored in a replay memory, from which mini-batches are randomly sampled during optimization. Furthermore, a target network is maintained to stabilize learning. The target Q-value is computed as

$$y_t = r_t + \gamma \max_{a'} Q_{\text{target}}(s_{t+1}, a'), \quad (12)$$

while the network parameters are optimized by minimizing the Bellman loss.

$$L(\theta) = \frac{1}{N} \sum_{i=1}^N (y_i - Q(s_i, a_i; \theta))^2, \quad (13)$$

The target network parameters are periodically synchronized with the online network, thereby reducing oscillations during training and improving convergence stability. Algorithm 1 summarizes the learning and evaluation procedures. Each training episode samples one reference lesion component and one axial slice intersecting that component; the environment exposes only the current ROI patch and two-dimensional state variables to the agent, while the selected reference box is used to compute the training reward. For internal evaluation, the reference mask is again processed to enumerate eligible components and select target-intersecting slices before the frozen policy is executed. One greedy episode is run for each reference-enumerated component and one prediction is matched to that assigned target. Thus, ground truth is not a

DQN input but is required by the evaluation harness. The protocol evaluates refinement of known lesion targets and cannot estimate how many lesions are present in an unseen volume.

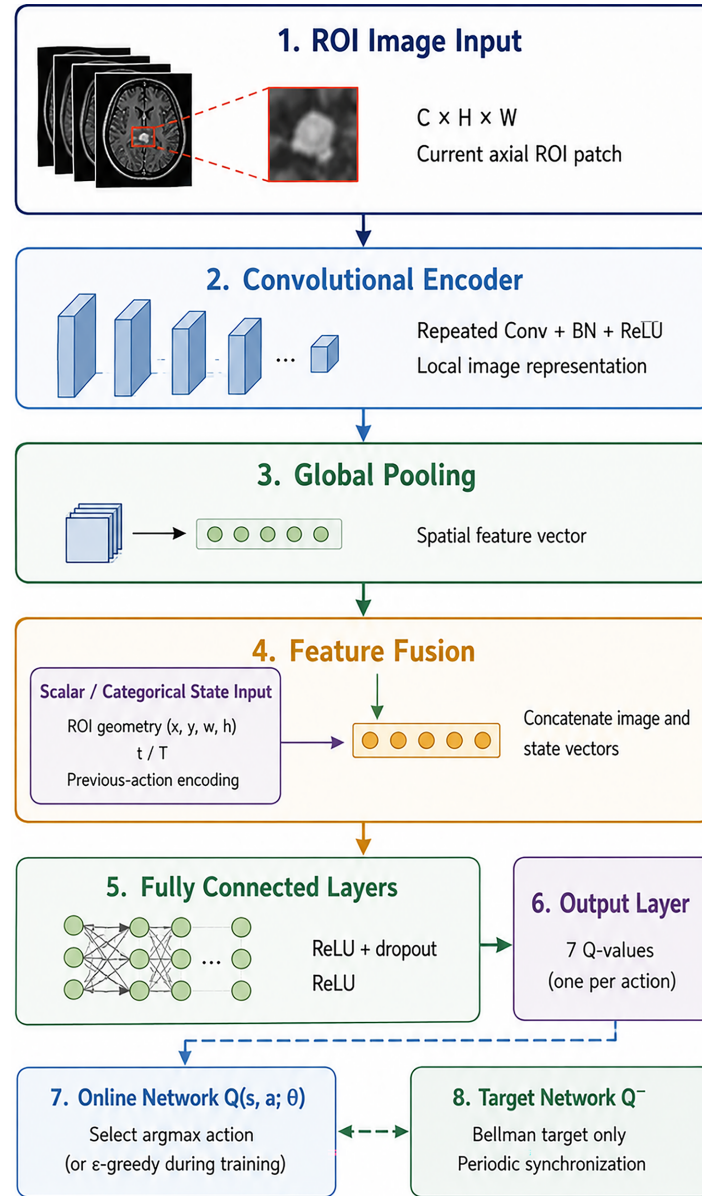


Figure 3: DQN architecture for two-dimensional multi-channel ROI localization without reference-mask state input.

Algorithm 1: SBL-DQN training and oracle-conditioned lesion-wise evaluation

Input:
 Axial multi-channel MRI slice I
 Reference lesion-component annotation G (training reward; evaluation candidate enumeration and scoring)
 Online Deep Q Network $Q(s, a; \theta)$
 Target Deep Q Network $Q_{\text{target}}(s, a; \theta^-)$

(Continued)

Algorithm 1 (continued)

Replay memory D
 Action set $A = \{\text{up, down, left, right, zoom in, zoom out, stop}\}$
 Discount factor γ
 Exploration rate ϵ
 Maximum number of steps T

```

1:      Initialize online DQN parameters  $\theta$  randomly
2:      Initialize target DQN parameters  $\theta- \leftarrow \theta$ 
3:      Initialize replay memory D
4:      for each training episode do
5:        Load an axial multi-channel MRI slice I
6:        Apply preprocessing to I
7:        Initialize the 2D search region B0 over the brain field
8:        Construct s0 from the 2D ROI patch, center (x, y), size (w, h), normalized step index, and
        previous-action code
9:        for t = 0 to T do
10:         With probability  $\epsilon$ , select a random action  $a_t \in A$ 
11:         Otherwise, select  $a_t = \operatorname{argmax}_a Q(s_t, a; \theta)$ 
12:         Execute action  $a_t$  in the MRI environment
13:         Update the search region Bt according to the selected action
14:         Construct next state  $s_{t+1}$ 
15:         Compute  $r_t$  from IoU, 2D centroid distance, area mismatch, step penalty, and
        Stop criterion
16:         if  $a_t = \text{stop}$  or maximum localization accuracy is achieved then
17:           Set donet = 1
18:         else
19:           Set donet = 0
20:         end if
21:         Store transition  $(s_t, a_t, r_t, s_{t+1}, \text{donet})$  in replay memory D
22:         Sample a random mini-batch of transitions from D
23:         Compute target value:
             $y_t = r_t + \gamma \max_{a'} Q_{\text{target}}(s_{t+1}, a'; \theta-)$ , if donet = 0
             $y_t = r_t$ , if donet = 1
24:         Update  $\theta$  by minimizing:  $L(\theta) = (y_t - Q(s_t, a_t; \theta))^2$ 
25:         Every C steps, update target network:  $\theta- \leftarrow \theta$ 
26:         Set  $s_t \leftarrow s_{t+1}$ 
27:         if donet = 1 then
28:           break
29:         end if
30:       end for
31:       Decay exploration rate  $\epsilon$ 
32:     end for
33:     Load the evaluation reference mask and enumerate eligible 26-connected components
    (oracle step)
  
```

(Continued)

Algorithm 1 (continued)

-
- | | |
|-----|--|
| 34: | For each reference component, select an intersecting axial slice, initialize the brain-field ROI, and apply $a = \text{argmax}_a Q(st, a; \theta)$ |
| 35: | Stop when the stop action is selected or T is reached; compute reward trace and metrics against the assigned reference target |
| 36: | Return one lesion-wise $B^* = (x, y, w, h)$, $C^* = (x, y)$, and P^* per reference-enumerated component |
-

[Table 1](#) consolidates the DQN, environment, reward, timing, and candidate-source settings used in the reported protocol. Candidate enumeration from reference masks is completed before the timed DQN search; the 0.18 s value therefore excludes preprocessing and oracle target construction.

4 Dataset

Three public collections were used. BraTS 2025 adult glioma [25] was the development source for training and internal evaluation. BraTS-METS 2025 [26] and QIN-GBM-TREATMENT-RESPONSE [27] were used for qualitative transfer examples and were not used for hyperparameter selection. This wording distinguishes the internal quantitative evidence from the external qualitative evidence and avoids an unsupported claim of strong cross-dataset generalization.

[Fig. 4](#) illustrates representative axial MRI samples and the construction of lesion localization targets across the three datasets employed in this study. The BraTS 2025 examples demonstrate complementary T1ce and FLAIR appearances of adult glioma, with expert voxel-level annotations converted into lesion contours, two-dimensional bounding boxes, and centroid targets. BraTS-METS 2025 presents a more complex multi-lesion scenario in which spatially separated brain metastases are represented as distinct connected components, allowing an individual bounding box and centroid to be defined for each lesion. The QIN-GBM-TREATMENT-RESPONSE example further demonstrates the variability of lesion appearance under a different MRI acquisition setting. Across the datasets, green contours denote reference lesion annotations, yellow rectangles represent the corresponding localization targets, and red points identify lesion centroids. Thus, [Fig. 4](#) visually demonstrates how heterogeneous voxel-level or reference-region annotations are transformed into standardized lesion-wise localization targets while preserving differences in MRI modality, lesion morphology, size, and multiplicity across the evaluated datasets.

[Table 2](#) summarizes the three public MRI collections, their imaging modalities, cohort characteristics, lesion types, annotation strategies, spatial resolution, and specific roles within the present evaluation protocol. BraTS 2025 Task 1 contains 2877 multi-institutional pre- and post-treatment adult glioma cases with co-registered T1, T1ce, T2, and FLAIR images and voxel-level labels [25]. For this localization study, the whole-tumor union of nonzero tumor labels was used. Three-dimensional 26-connected components were generated directly from the reference masks after the patient-level split; each intersecting two-dimensional component was converted to a minimum enclosing 2D bounding box and centroid. This reference-derived procedure was used during internal testing and constitutes oracle candidate enumeration. Patient identifiers, not slices or components, were partitioned before augmentation to prevent leakage.

BraTS-METS 2025 includes 1778 pre- and post-treatment brain-metastasis cases (1296 training, 179 validation, and 303 hidden testing cases in the official challenge organization) and requires T1, post-contrast T1, and FLAIR, with T2 optional [26]. Individual enhancing-tumor connected components were treated as distinct targets. QIN-GBM-TREATMENT-RESPONSE contains multiparametric MRI from 54 patients across 105 acquisition time points [27].

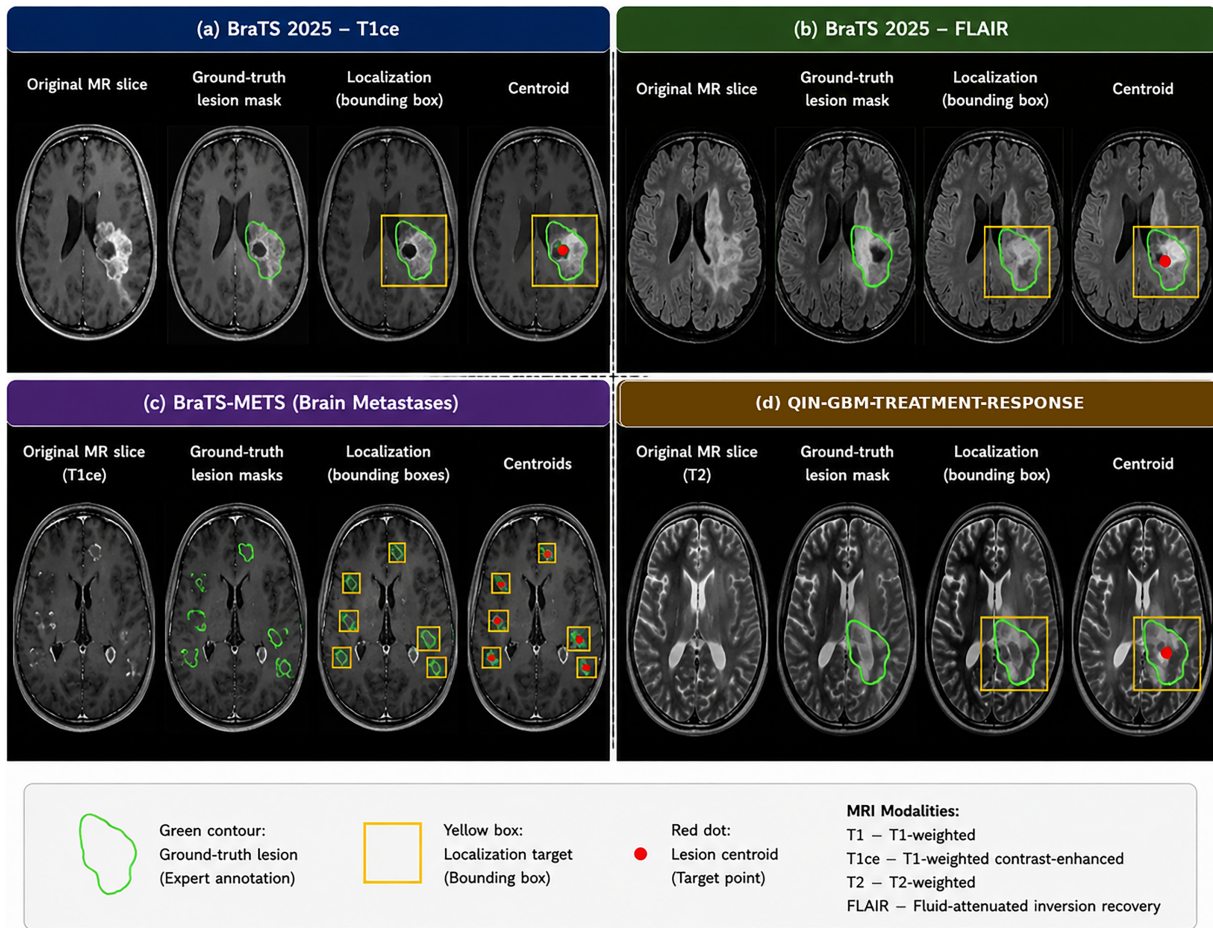


Figure 4: Representative axial MRI localization targets from BraTS 2025, BraTS-METS 2025, and QIN-GBM-TREATMENT-RESPONSE.

Table 2: Public MRI collections and their role in the present study.

Dataset	Role	MRI Modalities	Official Cohort	Lesion Type	Target Annotation	Scale
BraTS 2025 [25]	Development; internal metrics	T1, T1ce, T2, FLAIR	2877 cases	Pre/post-treatment adult glioma	Reference-mask 2D boxes	1 mm ³
BraTS-METS 2025 [26]	External qualitative; no tuning	T1, T1ce, FLAIR; T2 optional	1778 cases	Brain metastases	Reference-mask episodes	Native/1 mm ³
QIN-GBM-TREATMENT-RESPONSE [27]	External qualitative; no tuning	Structural + perfusion MRI	54 patients; 105 time points	Newly diagnosed glioblastoma	Reference-ROI episode	Varies

A small lesion was operationalized as a connected target with maximum in-plane diameter ≤ 10 mm after resampling. Eligible targets were identified from the reference-mask connected-component manifest after resampling. Each positive episode contains one assigned component and one intersecting axial slice; other disconnected components are excluded from that episode's reward target. Multi-lesion volumes are evaluated by running one episode per reference-enumerated component and matching one prediction to the assigned component.

Table 3 summarizes the study-specific patient-level partitioning and evidence roles across the three datasets, with BraTS 2025 divided into 2014 training (70%), 432 validation (15%), and 431 internal test subjects (15%) before oracle candidate enumeration, thereby preventing patient-level leakage. In contrast, BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE were excluded from model training and hyperparameter tuning and were used only to provide selected qualitative external examples. For BraTS-METS 2025, evaluation was conducted through lesion-wise target-conditioned episodes, whereas QIN-GBM-TREATMENT-RESPONSE was similarly examined without adaptation or parameter optimization. This partitioning strategy clearly separates model-development evidence from external qualitative evidence and ensures that the reported internal quantitative metrics originate exclusively from the held-out BraTS 2025 test partition.

Table 3: Study-specific patient-level data partition and evidence role.

Dataset	Training Subjects	Validation Subjects	Internal Test/External Cases	Partition Rule	Role
BraTS 2025	2014 (70%)	432 (15%)	431 (15%)	Patient-wise; oracle after split	Model development and internal metrics
BraTS-METS 2025	0	0	Selected external cases	No tuning; lesion-wise target-conditioned episodes	Qualitative external examples
QIN-GBM-TREATMENT-RESPONSE	0	0	Selected external cases	No tuning	Qualitative external examples

5 Results

The evaluation is organized around the evidence available for the oracle-conditioned lesion-wise two-dimensional protocol. Quantitative localization, convergence, and computational results are reported for the internal BraTS 2025 evaluation. Selected BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE cases are qualitative only. Reinforcement-learning baselines are compared under the same reference-mask candidate-episode protocol, whereas segmentation studies are retained only for task positioning because their voxel-mask outputs and metrics are not directly commensurate.

5.1 Evaluation Metrics

The evaluation unit is one oracle-conditioned axial candidate episode. A positive episode is assigned one reference lesion component; a background-only episode is negative. A prediction is a true positive when its final box satisfies both $\text{IoU} \geq 0.50$ and $\text{CLE} \leq 5$ mm with the assigned lesion. A positive episode that fails either criterion is a false negative. A terminated box on a background episode, or an unmatched duplicate box, is a false positive. A background episode that ends without a box above the confidence threshold $P^* = 0.50$ is a true negative. One-to-one matching prevents a reference component from generating more than one true positive. These are conditional episode-level refinement statistics; they are not examination-level sensitivity or specificity for autonomous lesion discovery.

Accuracy, Precision, Recall, and F1-score were computed from the episode-level TP, TN, FP, and FN definitions above [28]. Accuracy is

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (14)$$

where TP , TN , FP , and FN are counts of correctly localized positive episodes, correctly rejected background episodes, false localizations, and missed positive episodes, respectively. *Precision* and *Recall* are

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (15)$$

whereas *Recall* measures the proportion of correctly localized lesions among all actual lesions,

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (16)$$

To balance both precision and recall, the harmonic mean of these measures was calculated using the F1-score,

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (17)$$

These metrics collectively evaluate the capability of the proposed SBL-DQN framework to accurately identify lesion locations while minimizing both missed detections and false localization results.

Localization geometry was assessed with two-dimensional Intersection over Union (IoU) and Center Localization Error (CLE) [29]. IoU is the overlap between predicted box B_p and ground-truth box B_g box on the axial plane,

$$\text{IoU} = \frac{|B_p \cap B_g|}{|B_p \cup B_g|}. \quad (18)$$

Higher IoU indicates better in-plane overlap. CLE is the Euclidean distance between predicted centroid C_p and reference centroid C_g in millimeters,

$$\text{CLE} = \sqrt{[(x(\text{pred}) - x_g))^2 + (y(\text{pred}) - y_g))^2}, \quad (19)$$

where (x_p, y_p) and (x_g, y_g) are the predicted and reference in-plane centroid coordinates after conversion to millimeters. Average precision is reported at a single IoU threshold of 0.50, not as mean AP across thresholds,

$$\text{AP@0.5} = \int_0^1 \hat{p}(r) dr, \quad (20)$$

where $\tilde{p}(r)$ is interpolated precision at recall r after sorting predictions by confidence. The final confidence is $P^* = \exp(Q(s_T, \text{Stop})) / \sum_a \exp(Q(s_T, a))$, i.e., the softmax-normalized terminal Stop Q-value.

Since SBL-DQN is based on reinforcement learning, several reinforcement learning-specific evaluation measures were additionally employed. The Episode Success Rate (ESR) quantifies the proportion of localization episodes that successfully reached the lesion before the maximum episode length,

$$ESR = \frac{N_{success}}{N_{episodes}} \times 100\% \quad (21)$$

The Average Episode Reward (AER) evaluates the overall effectiveness of the learned policy by averaging cumulative rewards over all testing episodes [30],

$$AER = \frac{1}{N_{episodes}} \sum_{i=1}^{N_{episodes}} R_i, \quad (22)$$

where R_i denotes the cumulative reward obtained during the i -th episode. Furthermore, the Average Search Steps (ASS) metric measures the average number of actions required to localize a lesion,

$$ASS = \frac{1}{N_{episodes}} \sum_{i=1}^{N_{episodes}} S_i, \quad (23)$$

where S_i represents the number of executed actions before episode termination. Lower ASS values indicate more efficient search strategies learned by the reinforcement learning agent.

Computational efficiency was assessed with search latency and model complexity. Search latency is the elapsed time required by the DQN to process a predefined volume-level candidate set after preprocessing and candidate enumeration; it is not end-to-end examination time. Model complexity is the number of trainable parameters.

$$T_{inf} = \frac{1}{N} \sum_{i=1}^N t_i, \quad (24)$$

where t_i is the search time for evaluation candidate set i and N is the number of timed candidate sets. The hardware, software, warm-up, and repetition details required to reproduce this measurement are reported in Table 1. Together, the metrics characterize the reported lesion-wise protocol; they do not establish clinical utility or broad external generalization.

5.2 Experiment Results

Quantitative convergence, localization, search-efficiency, and computational results were obtained on the internal BraTS 2025 oracle-conditioned evaluation. BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE are represented by selected qualitative examples only; no dataset-level external metric table is claimed. Comparisons with Vanilla DQN, Double DQN, Dueling DQN, and PPO use the same reference-enumerated lesion-wise protocol. Segmentation methods are discussed only to position the output task and are not treated as commensurate baselines.

Fig. 5 illustrates the evolution of average episode return for SBL-DQN and four baseline methods across 1800 training episodes. All approaches begin with strongly negative returns, indicating inefficient exploratory actions, limited localization accuracy, and frequent accumulation of distance, size, and step penalties during early training. SBL-DQN improves most rapidly, crosses the zero-return threshold at approximately 150 episodes, and reaches about 40 by 500 episodes. Its return subsequently increases more gradually, stabilizing

around 50–55 after roughly 1000 episodes, with only small oscillations thereafter. Double DQN exhibits the second-fastest improvement, becomes positive near 300 episodes, and converges around 20–25. Dueling DQN crosses zero later, at approximately 400 episodes, before plateauing near 15. PPO requires about 700–750 episodes to obtain positive returns and ultimately stabilizes between 8 and 10. Vanilla DQN improves steadily but remains slightly negative at the end of training, suggesting less efficient search behavior and weaker terminal localization performance. The widening separation between SBL-DQN and the baselines during the first 800 episodes demonstrates superior learning efficiency, while the flatter trajectories after approximately 1200 episodes indicate convergence. Moreover, the limited late-stage fluctuations suggest stable policy updates rather than reward divergence. Overall, SBL-DQN achieves the highest final return and the most favorable balance between localization quality, trajectory efficiency, and successful stopping behavior under the common 300-action protocol for small brain lesion localization.

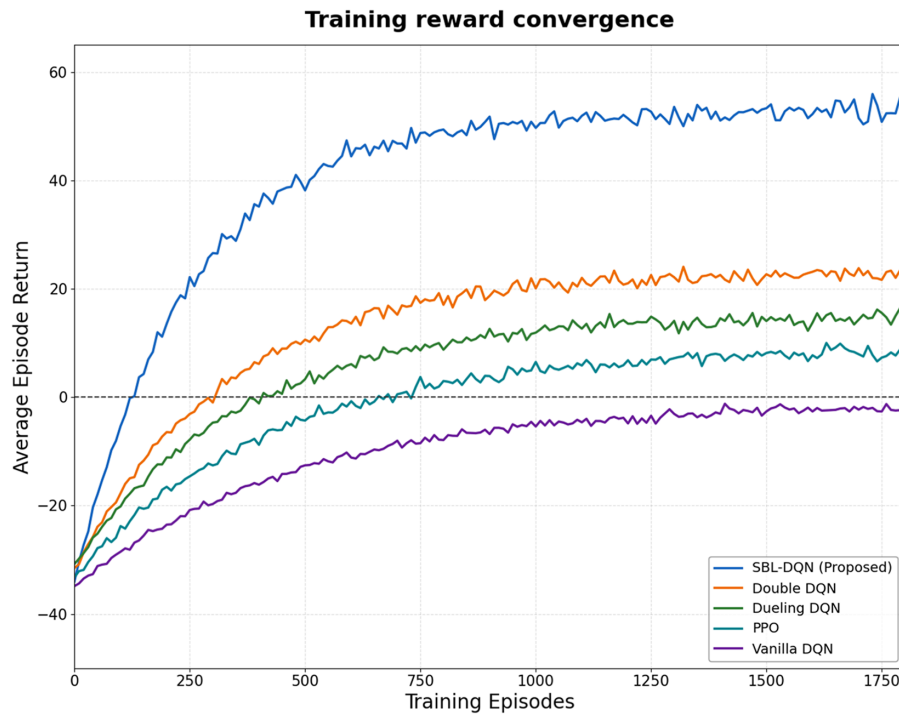


Figure 5: Training reward convergence of the proposed SBL-DQN and baseline reinforcement learning methods.

Fig. 6 demonstrates the progressive improvement in localization accuracy achieved by SBL-DQN and the four baseline reinforcement learning methods on the BraTS 2025 validation set as training proceeds. The proposed SBL-DQN exhibits the fastest convergence, increasing from approximately 7% during the initial stage to 50% at 100 episodes and approximately 69% at 280 episodes. Its localization accuracy subsequently reaches approximately 85% at 600 episodes and approaches 90% by 800 episodes. Continued training produces smaller but consistent improvements, with accuracy exceeding 91% at 1000 episodes and ultimately reaching approximately 94.5% after 1800 episodes. Double DQN represents the strongest baseline, attaining approximately 86.5% at the end of training, corresponding to an approximately eight-percentage-point difference from SBL-DQN. Dueling DQN converges at approximately 78%, while PPO reaches approximately 65.7%. Vanilla DQN exhibits the weakest localization performance, plateauing near 47.5% despite continued training. Notably, the performance curves progressively flatten after approximately 1000–1200 episodes, indicating convergence toward stable localization policies. The consistently higher SBL-DQN trajectory

across all training stages demonstrates superior learning efficiency and final localization performance under the common validation protocol, although these results characterize the complete configuration rather than the isolated effects of individual model components.

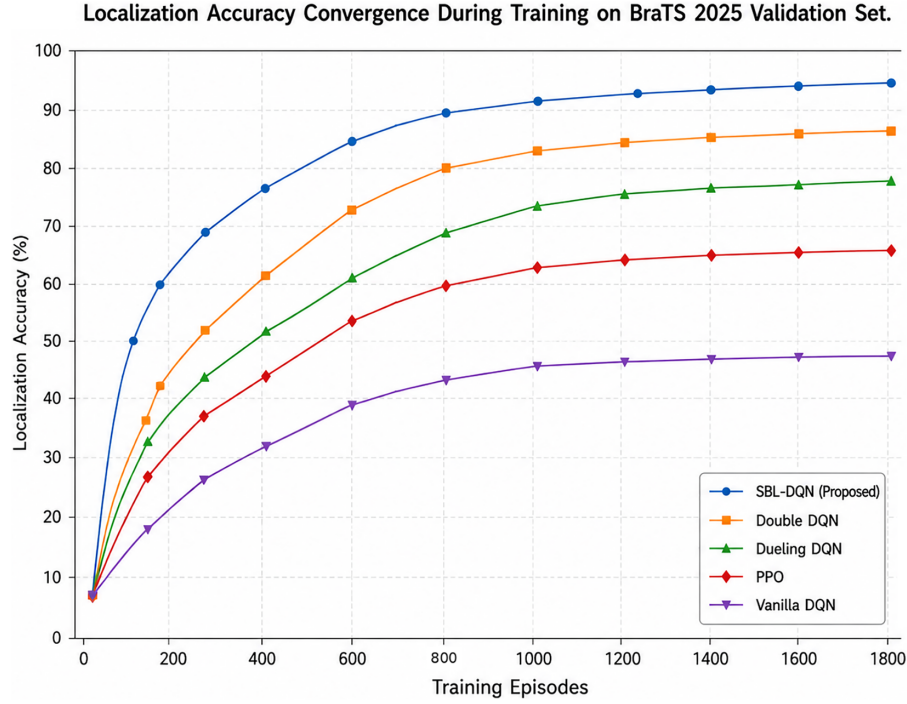


Figure 6: Localization accuracy convergence of the proposed SBL-DQN and baseline reinforcement learning methods on the BraTS 2025 validation set.

Fig. 7 demonstrates the evolution of mean Intersection-over-Union (IoU) during training on the BraTS 2025 validation set, revealing substantial differences in localization quality among the evaluated reinforcement learning methods. The proposed SBL-DQN exhibits the fastest initial improvement, achieving a mean IoU of approximately 0.38 after 100 episodes and 0.53 after 200 episodes. Performance continues to increase steadily, reaching approximately 0.74 at 600 episodes and 0.80 at approximately 1100 episodes. Beyond this stage, the curve gradually approaches a stable plateau, ultimately attaining a mean IoU of approximately 0.83. Double DQN provides the strongest baseline performance, converging near 0.73, approximately 0.10 below SBL-DQN. Dueling DQN reaches approximately 0.61, whereas PPO stabilizes near 0.49. Vanilla DQN demonstrates substantially weaker spatial localization, with its mean IoU plateauing at approximately 0.33. The progressive flattening of all curves after approximately 1200–1400 episodes indicates that additional training produces increasingly marginal improvements. Importantly, SBL-DQN maintains the highest mean IoU throughout the training process, demonstrating more accurate overlap between predicted and reference lesion regions under the common validation protocol. Nevertheless, the observed advantage represents the performance of the complete SBL-DQN configuration and cannot independently establish the contribution of its state representation, zoom operations, or individual reward components.

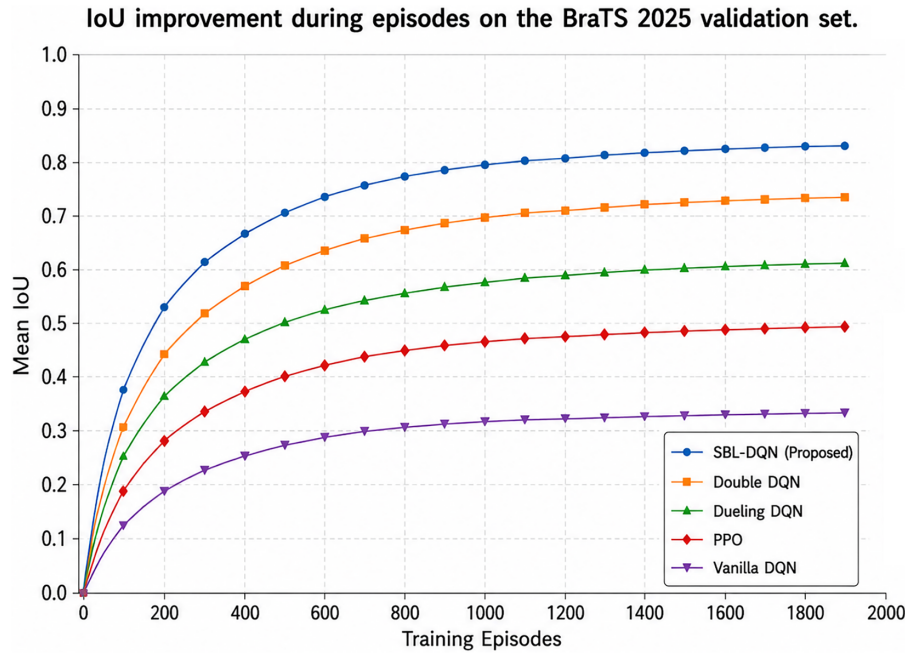


Figure 7: Mean intersection-over-union (IoU) convergence during training on the BraTS 2025 validation set.

Fig. 8 presents representative search trajectories generated by the proposed SBL-DQN for three lesion examples drawn from BraTS 2025, BraTS-METS 2025, and QIN-GBM-TREATMENT-RESPONSE, providing qualitative evidence of the learned sequential localization behavior. In each example, the agent begins with a relatively broad region of interest and progressively adjusts its spatial position and dimensions through successive actions. The intermediate steps demonstrate simultaneous translation toward the target and contraction of the search region, ultimately producing a predicted localization closely aligned with the ground-truth lesion. The corresponding trajectory plots further characterize this process in terms of the x- and y-coordinates and ROI area. Importantly, the third axis represents the changing two-dimensional ROI area rather than anatomical depth or movement between MRI slices. Across the three examples, ROI area generally decreases as the trajectory approaches the final localization, indicating progressive spatial refinement rather than exhaustive exploration of the image. The examples additionally suggest that the learned policy can accommodate differences in lesion position, appearance, and size across glioma, metastatic, and glioblastoma cases. Nevertheless, the BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE trajectories constitute selected qualitative examples. Consequently, they demonstrate the feasibility of transferring the learned search behavior to heterogeneous MRI cases but should not be interpreted as quantitative evidence of cross-dataset generalization.

Fig. 9 presents selected successful examples and Grad-CAM maps. Grad-CAM highlights image regions whose final-layer activations are associated with the selected action value [31]; it does not prove clinical interpretability, causal reasoning, or clinical relevance. The shown maps overlap the visible lesion in these examples, but no reader study or attribution-faithfulness analysis was performed. These cases should therefore be interpreted as model-inspection aids rather than clinical validation.

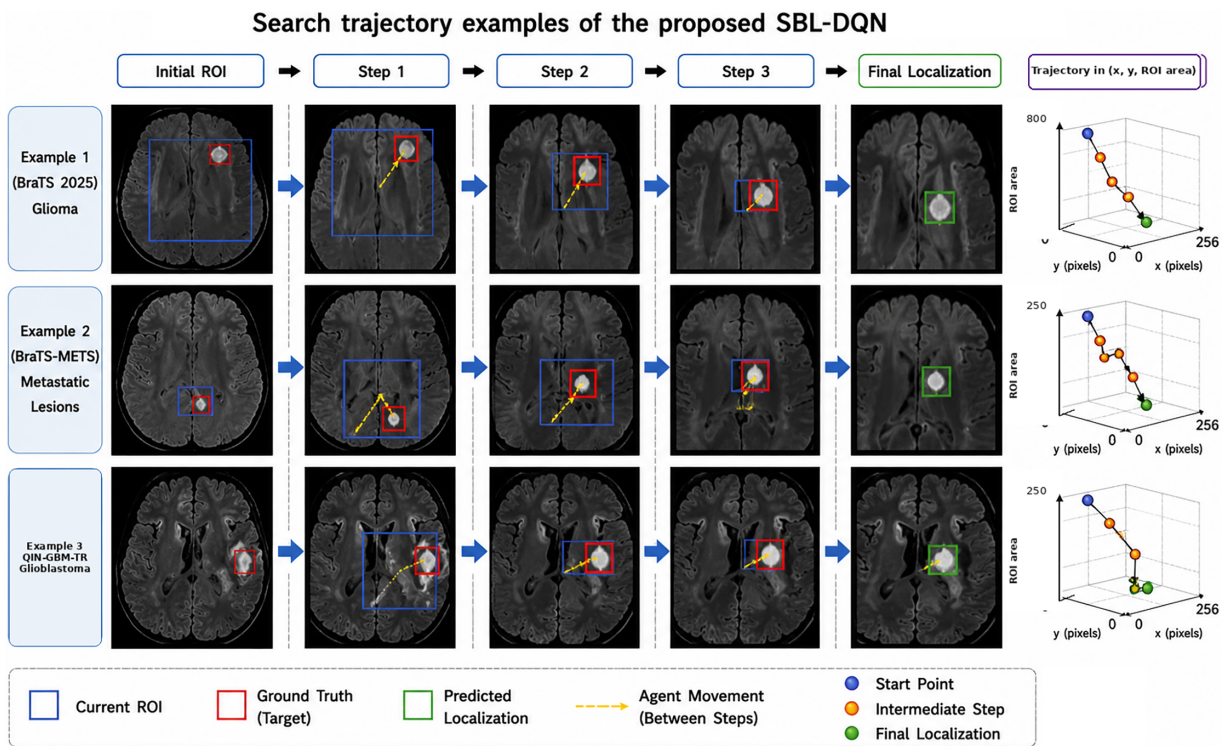


Figure 8: Representative two-dimensional search trajectories; the third axis denotes ROI area rather than anatomical depth.

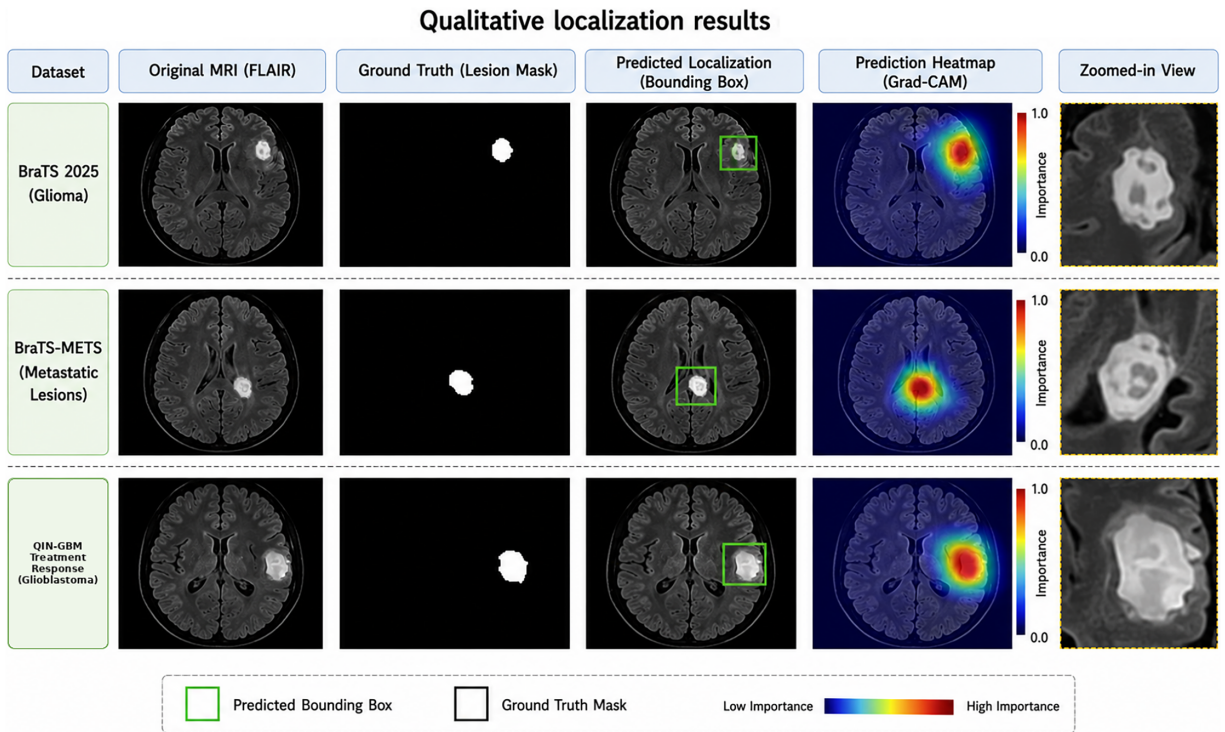


Figure 9: Illustrative localization and Grad-CAM maps on BraTS 2025, BraTS-METS 2025, and QIN-GBM-TREATMENT-RESPONSE.

5.3 Component-Wise Ablation and Failure-Case Evidence

The retained experiment record contains the complete SBL-DQN configuration and algorithm-level baselines, but no controlled component-wise reruns. The baseline curves cannot be re-labeled as ablations because they change the reinforcement-learning algorithm rather than isolating the proposed state representation, adaptive zoom strategy, or composite reward. A valid ablation must hold the patient split, initialization, training budget, random seeds, and evaluation candidate list fixed while comparing at minimum: (i) the full model; (ii) a state variant without ROI geometry, episode progress, and action history; (iii) a translation-and-stop variant without zoom actions; and (iv) an IoU-only reward variant. Until those runs are performed, the integrated results do not establish the independent contribution of any one component.

The available prediction archive contains selected successful cases only and does not retain the case-level boxes, masks, confidence scores, or residuals needed to sample false-positive, false-negative, low-IoU, and large-center-error examples. For a balanced audit, false positive is defined as a confident localization on a background episode or an unmatched duplicate; false negative is an assigned positive episode failing $\text{IoU} \geq 0.50$ or $\text{CLE} \leq 5$ mm; low IoU is $0 < \text{IoU} < 0.50$; and large center error is $\text{CLE} > 5$ mm. A complete qualitative panel should sample equal numbers from these four categories together with successful controls from the locked internal test set. Fig. 9 is therefore identified only as a selected success panel. No missing cases or per-dataset counts have been manufactured, and claims of comprehensive robustness are withdrawn pending regeneration of the component manifest and per-case prediction archive.

Fig. 10 compares the complete SBL-DQN configuration with four reinforcement learning baselines. Fig. 10a reports $\text{AP@0.5} = 0.961$ for SBL-DQN, followed by 0.913, 0.842, 0.741, and 0.576 for Double DQN, Dueling DQN, PPO, and Vanilla DQN. Fig. 10b reports final CLE near 0.9 mm, and Fig. 10c shows episode success approaching 95% at the 300-step limit. These values are internal results under the common lesion-wise protocol.

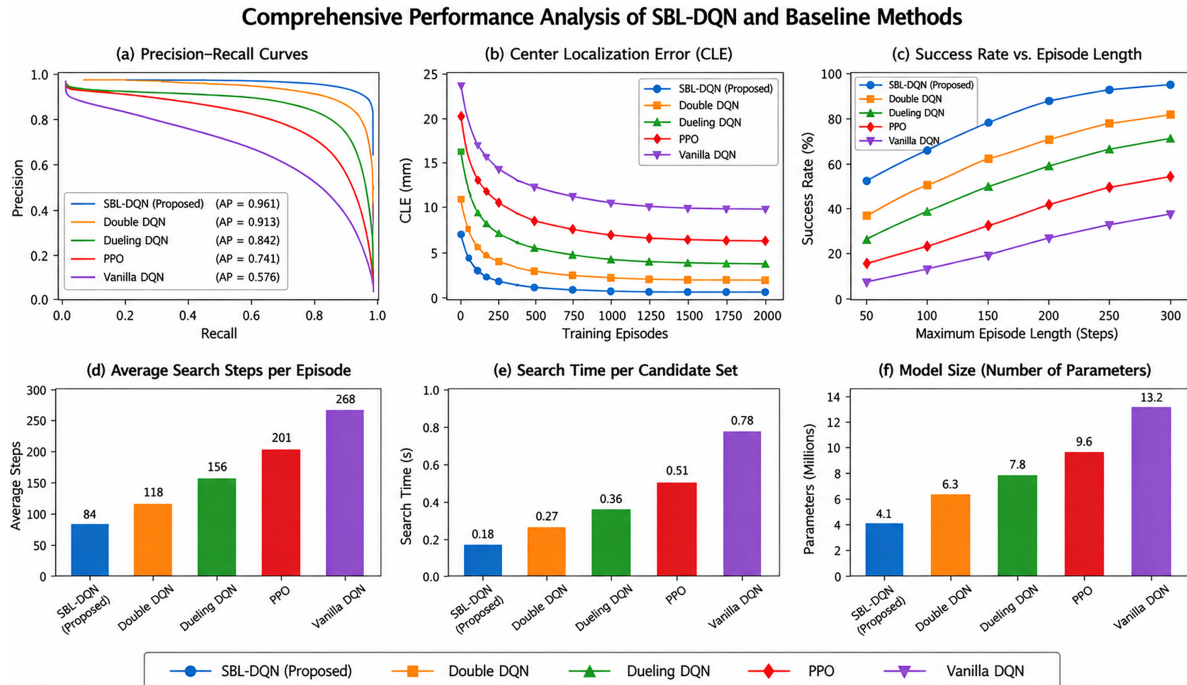


Figure 10: Comprehensive performance comparison of the proposed SBL-DQN and baseline reinforcement learning methods.

Fig. 10d–f reports 84 average search steps, 0.18 s search time per predefined volume-level candidate set, and 4.1 million trainable parameters for SBL-DQN. The 0.18 s value excludes image loading, registration, resampling, normalization, candidate enumeration, and visualization; it was measured with batch size 1 over 100 BraTS 2025 internal-test volumes on an NVIDIA RTX 3090 24 GB GPU with an Intel Xeon Gold 6226R CPU, using PyTorch 2.1.0 and CUDA 12.1 after 10 warm-up runs. It is therefore described as search latency rather than real-time or end-to-end clinical performance.

Table 4 positions SBL-DQN by task and output rather than combining non-commensurate numerical metrics. Stember and Shalu [9] are the closest direct precedent because they also use DQN/Q-learning for 2D brain tumor localization. Dense segmentation methods predict voxel masks and are therefore not head-to-head baselines for the sequential bounding-box task. No superiority over those segmentation methods is inferred from the comparison.

Table 4: Task-oriented positioning of SBL-DQN relative to medical-Image localization and segmentation methods.

Study	Imaging Target	Search/Output Formulation	Evidence Reported	Relationship to SBL-DQN
Ghesu et al. [10]	CT anatomical landmarks	Multi-scale 3D RL; landmark point	Same-study landmark accuracy and speed	Sequential multi-scale localization precedent; different anatomy and output
Alansary et al. [21]	Anatomical landmarks in 3D medical scans	DRL navigation; landmark point	Agent evaluation on landmark tasks	RL localization precedent; not lesion bounding boxes
Maicas et al. [22]	Breast lesions in DCE-MRI	DQN attention/search; lesion detection	Detection accuracy and runtime on 117 volumes	Lesion-search precedent in MRI; different organ and 3D formulation
Stember and Shalu [9]	One brain tumor per 2D T1ce slice	Three-action 60 × 60 gridworld; lesion-overlapping cell	30 train/30 test slices; 70% test accuracy	Direct DQN brain-tumor precedent; SBL-DQN changes state, actions, reward, and box output
Swin UNETR [32]/nnU-Net [33]	Brain-tumor segmentation	Dense voxel-wise masks	Dice/surface metrics on BraTS benchmarks	Context only; outputs and metrics are not directly commensurate
Proposed SBL-DQN	Small lesion candidate on axial brain MRI	Seven-action 2D ROI refinement; box, centroid, P*	Internal BraTS 2025: accuracy 97.46%, IoU 0.834, AP@0.5 0.961, CLE 0.91 mm	Lesion-wise target-conditioned localization; no claim of superiority over segmentation

Overall, the internal experiments show that the complete SBL-DQN configuration converges faster and yields better oracle-conditioned lesion-wise localization metrics than the evaluated reinforcement-learning baselines. The present evidence does not isolate the effects of the state representation, zoom actions, or reward

terms; it does not quantify autonomous candidate generation; and it does not provide dataset-level external validation or a balanced failure taxonomy. The conclusions are restricted accordingly.

6 Discussion

The internal results support sequential two-dimensional ROI refinement for a reference-defined lesion target. SBL-DQN achieved higher reward, localization accuracy, and IoU than four reinforcement-learning baselines under the same oracle-conditioned protocol. These observations apply only to the integrated configuration. Because controlled component-wise runs are absent, the differences cannot be assigned independently to the expanded state, zoom actions, or composite reward.

The experiment is explicitly oracle-conditioned. Ground-truth masks are processed at test time to enumerate eligible connected components and select target-intersecting axial slices; one episode is then constructed for each assigned component. Although the agent state excludes the reference mask and target coordinates, the evaluation cannot be initiated without these annotations. The method therefore does not infer how many lesions are present in an unseen volume and is not an autonomous multiple-metastasis detector. A separately validated image-only candidate generator, duplicate suppression, and volume-level matching are required before autonomous inference can be claimed.

The external figures show that the trained policy can produce plausible boxes on selected BraTS-METS 2025 and QIN-GBM-TREATMENT-RESPONSE examples. Because dataset-specific quantitative estimates, confidence intervals, and balanced failure cases are not available, these examples are evidence of feasibility rather than strong generalization. Grad-CAM maps likewise indicate gradient-associated regions only and do not establish clinical interpretability or relevance [31].

Additional limitations include the two-dimensional axial formulation, oracle target enumeration, unavailable eligible-lesion counts, missing controlled component-wise ablation, absence of a retained balanced failure-case archive, and search-only rather than end-to-end timing. Future work should operate directly on three-dimensional volumes, discover and match multiple lesions without ground-truth enumeration, regenerate per-dataset component manifests, run fixed-protocol ablations across independent seeds, report confidence intervals, and include systematic failure cases before clinical translation is considered.

7 Conclusion

This study presented SBL-DQN, a two-dimensional deep reinforcement learning framework for iterative bounding-box refinement of reference-defined lesion components on axial multi-channel MRI. The agent operates through four translation actions, two scale-adjustment actions, and a termination action to progressively localize each assigned target. Under the oracle-conditioned BraTS 2025 internal evaluation protocol, SBL-DQN achieved 97.46% localization accuracy, 0.834 mean IoU, 0.961 AP@0.5, and 0.91 mm center localization error, outperforming the evaluated reinforcement-learning baselines. The framework integrates ROI-based image information, spatial state variables, adaptive action selection, experience replay, and composite reward optimization. Nevertheless, these results should be interpreted within the defined experimental setting. Reference masks are used to enumerate lesion candidates before individual localization episodes; therefore, the reported performance represents oracle-conditioned lesion-wise refinement rather than autonomous lesion discovery. Moreover, the independent contributions of the state representation, action strategy, and reward components require controlled component-wise ablation. External-dataset results currently provide qualitative evidence only, while comprehensive eligible-lesion counts, systematic failure analysis, and end-to-end computational profiling remain necessary. Future work will consequently focus on autonomous candidate generation, rigorous ablation studies, quantitative multi-center external validation, failure-mode characterization, and complete end-to-end efficiency assessment to establish the framework's generalizability and practical applicability.

Acknowledgement: Not applicable.

Funding Statement: This study was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the framework of grant AP23489899 “Applying Deep Learning and Neuroimaging Methods for Brain Stroke Diagnosis”.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Bakhytzhan Omarov and Balnur Kenjayeva; methodology, Bakhytzhan Omarov and Daniyar Sultan; software, Bakhytzhan Omarov and Zhanseri Ikram; validation, Bakhytzhan Omarov, Balnur Kenjayeva, Daniyar Sultan, and Zhanseri Ikram; formal analysis, Bakhytzhan Omarov and Daniyar Sultan; investigation, Bakhytzhan Omarov and Balnur Kenjayeva; resources, Balnur Kenjayeva; data curation, Bakhytzhan Omarov and Zhanseri Ikram; writing—original draft preparation, Bakhytzhan Omarov; writing—review and editing, Balnur Kenjayeva, Daniyar Sultan, and Zhanseri Ikram; visualization, Bakhytzhan Omarov and Zhanseri Ikram; supervision, Balnur Kenjayeva and Daniyar Sultan; project administration, Balnur Kenjayeva; funding acquisition, Balnur Kenjayeva. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: BraTS 2025 and BraTS-METS 2025 are available through the BraTS 2025 Synapse challenge workspace (Synapse ID: syn64153130; access subject to registration and challenge terms). QIN-GBM-TREATMENT-RESPONSE is hosted by The Cancer Imaging Archive under DOI 10.7937/K9/TCIA.2016.nQF4gpn2.

Ethics Approval: This study used publicly available, de-identified cancer imaging data. No new human participant recruitment or data collection was conducted, and no identifiable patient information was accessed. The study was determined to be exempt from ethics approval.

Conflicts of Interest: Given his role as Guest Editor of this Special Issue, Bakhytzhan Omarov had no involvement in the peer review of this article and had no access to information regarding its peer review. Full responsibility for the editorial process for this article was delegated to another journal editor. The authors declare no other conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AER	Average Episode Reward
AP	Average Precision
ASS	Average Search Steps
BraTS	Brain Tumor Segmentation Challenge
CLE	Center Localization Error
CNN	Convolutional Neural Network
DCE-MRI	Dynamic Contrast-Enhanced Magnetic Resonance Imaging
DRL	Deep Reinforcement Learning
DQN	Deep Q Network
ESR	Episode Success Rate
FLAIR	Fluid-Attenuated Inversion Recovery
IoU	Intersection over Union
MDP	Markov Decision Process
MRI	Magnetic Resonance Imaging
PPO	Proximal Policy Optimization
QIN	Quantitative Imaging Network
ReLU	Rectified Linear Unit
RL	Reinforcement Learning
ROI	Region of Interest
SBL-DQN	Small Brain Lesion Deep Q Network

T1	T1-Weighted MRI
T1ce	Contrast-Enhanced T1-Weighted MRI
T2	T2-Weighted MRI
TCIA	The Cancer Imaging Archive

References

1. Khalighi S, Reddy K, Midya A, Pandav KB, Madabhushi A, Abedalthagafi M. Artificial intelligence in neuro-oncology: advances and challenges in brain tumor diagnosis, prognosis, and precision treatment. *NPJ Precis Oncol.* 2024;8(1):80. doi:10.1038/s41698-024-00575-0.
2. Rastogi D, Johri P, Tiwari V, Elngar AA. Multi-class classification of brain tumour magnetic resonance images using multi-branch network with inception block and five-fold cross validation deep learning framework. *Biomed Signal Process Control.* 2024;88:105602. doi:10.1016/j.bspc.2023.105602.
3. Rasool N, Bhat JI. A critical review on segmentation of glioma brain tumor and prediction of overall survival. *Arch Comput Meth Eng.* 2025;32(3):1525–69. doi:10.1007/s11831-024-10188-2.
4. Tosee ZR, Jalali-Zefrei F, Chaboki BG, Karimian P, Mohammadi-Vajari E, Souiri Z. Integration of conventional MRI and diffusion-weighted imaging for differential diagnosis of high-grade gliomas and metastatic brain tumors. *Curr Radiopharm.* 2026;19(2):100022. doi:10.1016/j.craph.2025.100022.
5. Omarov B. Deep learning in biomedical image and signal processing: a survey. *Comput Mater Contin.* 2025;85(2):2195–253. doi:10.32604/cmc.2025.064799.
6. Rahman MA, Masum MI, Hasib KM, Mridha MF, Alfarhood S, Safran M, et al. GliomaCNN: an effective lightweight CNN model in assessment of classifying brain tumor from magnetic resonance images using explainable AI. *Comput Model Eng Sci.* 2024;140(3):2425–48. doi:10.32604/cmcs.2024.050760.
7. Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal.* 2017;36:61–78. doi:10.1016/j.media.2016.10.004.
8. Sampa MB, Abdul Aziz NH, Rahman MS, Ab Aziz NA, Besar R, Ghazali AK. Reinforcement learning for medical image analysis: a systematic review of algorithms, engineering challenges, and clinical deployment. *Comput Assist Surg.* 2026;31(1):2597553. doi:10.1080/24699322.2025.2597553.
9. Stemmer JN, Shalu H. Reinforcement learning using deep Q networks and Q learning accurately localizes brain tumors on MRI with very small training sets. *BMC Med Imaging.* 2022;22(1):224. doi:10.1186/s12880-022-00919-x.
10. Ghesu FC, Georgescu B, Zheng Y, Grbic S, Maier A, Hornegger J, et al. Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans. *IEEE Trans Pattern Anal Mach Intell.* 2019;41(1):176–89. doi:10.1109/TPAMI.2017.2782687.
11. Abu Owida H, AlMahadin G, Al-Nabulsi JI, Turab N, Abuowaida S, Alshdaifat N. Automated classification of brain tumor-based magnetic resonance imaging using deep learning approach. *Int J Electr Comput Eng.* 2024;14(3):3150. doi:10.11591/ijece.v14i3.pp3150-3158.
12. Das S, Goswami RS. Review, limitations, and future prospects of neural network approaches for brain tumor classification. *Multimed Tools Appl.* 2024;83(15):45799–841. doi:10.1007/s11042-023-17215-7.
13. Ghosh S, Das S. Multi-scale morphology-aided deep medical image segmentation. *Eng Appl Artif Intell.* 2024;137:109047. doi:10.1016/j.engappai.2024.109047.
14. Baumgartner M, Jäger PF, Isensee F, Maier-Hein KH. nnDetection: a self-configuring method for medical object detection. In: *Medical image computing and computer assisted intervention—MICCAI 2021*. Cham, Switzerland: Springer International Publishing; 2021. p. 530–9. doi:10.1007/978-3-030-87240-3_51.
15. Ghosal P, Roy A, Agarwal R, Purkayastha K, Sharma AL, Kumar A. Compound attention embedded dual channel encoder-decoder for ms lesion segmentation from brain MRI. *Multimed Tools Appl.* 2025;84(26):31139–71. doi:10.1007/s11042-024-20416-3.

16. Abbas J, Soomro DB, Huang S, Liu L. DualAttendMed: a coarse-to-fine dual-stage attention framework for interpretable disease localization and classification. *Expert Syst Appl.* 2026;305:130886. doi:10.1016/j.eswa.2025.130886.
17. Aiya AJ, Wani N, Ramani M, Kumar A, Pant S, Kotecha K, et al. Optimized deep learning for brain tumor detection: a hybrid approach with attention mechanisms and clinical explainability. *Sci Rep.* 2025;15(1):31386. doi:10.1038/s41598-025-04591-3.
18. Chen YT, Ahmad N, Aurangzeb K. Enhancing 3D U-Net with residual and squeeze-and-excitation attention mechanisms for improved brain tumor segmentation in multimodal MRI. *Comput Model Eng Sci.* 2025;144(1):1197–224. doi:10.32604/cmescs.2025.066580.
19. Wang W, Chen C, Ding M, Yu H, Zha S, Li J. TransBTS: multimodal brain tumor segmentation using transformer. In: *Medical image computing and computer assisted intervention—MICCAI 2021*. Cham, Switzerland: Springer International Publishing; 2021. p. 109–19. doi:10.1007/978-3-030-87193-2_11.
20. Alshomrani F. Challenges and advances in classifying brain tumors: an overview of machine, deep learning, and hybrid approaches with future perspectives in medical imaging. *Curr Med Imaging.* 2025;21:e15734056365191. doi:10.2174/0115734056365191250602124819.
21. Alansary A, Oktay O, Li Y, Folgoc LL, Hou B, Vaillant G, et al. Evaluating reinforcement learning agents for anatomical landmark detection. *Med Image Anal.* 2019;53:156–64. doi:10.1016/j.media.2019.02.007.
22. Maicas G, Carneiro G, Bradley AP, Nascimento JC, Reid I. Deep reinforcement learning for active breast lesion detection from DCE-MRI. In: *Medical image computing and computer assisted intervention—MICCAI 2017*. Cham, Switzerland: Springer International Publishing; 2017. p. 665–73. doi:10.1007/978-3-319-66179-7_76.
23. Milletari F, Navab N, Ahmadi SA. V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: *Proceedings of the 2016 Fourth International Conference on 3D Vision (3DV)*; 2016 Oct 25–28; Stanford, CA, USA. p. 565–71. doi:10.1109/3DV.2016.79.
24. Khajuria R, Sarwar A. Active reinforcement learning based approach for localization of target ROI (region of interest) in cervical cell images. *Multimed Tools Appl.* 2025;84(17):18467–79. doi:10.1007/s11042-024-19416-0.
25. Brain Tumor Segmentation (BraTS) Cluster of Challenges. BraTS 2025 lighthouse challenge: adult glioma segmentation on pre- and post-treatment MRI. Synapse ID: syn64153130. 2025 [cited 2026 Aug 20]. Available from: <https://www.synapse.org/brats2025>.
26. Maleki N, Amiruddin R, Moawad AW, Yordanov N, Gkampenis A, Fehringer P, et al. Analysis of the MICCAI brain tumor segmentation—metastases (BraTS-METS) 2025 lighthouse challenge: brain metastasis segmentation on pre- and post-treatment MRI. *arXiv:2504.12527*. 2025.
27. Mamonov AB, Kalpathy-Cramer J. Data from QIN-GBM-TREATMENT-RESPONSE [Dataset]. The cancer imaging archive. 2016. doi:10.7937/K9/TCIA.2016.nQF4gpn2.
28. Padilla R, Passos WL, Dias TLB, Netto SL, da Silva EAB. A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics.* 2021;10(3):279. doi:10.3390/electronics10030279.
29. Rezaatofghi H, Tsoi N, Gwak J, Sadeghian A, Reid I, Savarese S. Generalized intersection over union: a metric and a loss for bounding box regression. In: *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; 2019 Jun 15–20; Long Beach, CA, USA. p. 658–66. doi:10.1109/CVPR.2019.00075.
30. Zhang Y, Ross K. On-policy deep reinforcement learning for the average-reward criterion. In: *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*; 2021 Jul 18–24; Online. p. 5925–35.
31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*; 2017 Oct 22–19; Venice, Italy. p. 12535–45.
32. Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D. Swin UNETR: swin transformers for semantic segmentation of brain tumors in MRI images. In: *International MICCAI brainlesion workshop*. Cham, Switzerland: Springer International Publishing; 2021. p. 272–84. doi:10.1007/978-3-031-08999-2_22.
33. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Meth.* 2021;18(2):203–11. doi:10.1038/s41592-020-01008-z.