Missing Modality-Aware Calibration for Trustworthy Brain Tumor Segmentation

Sol Lee, Hyunji Kim, Sungrae Hong, Donghee Han, and Mun Yong Yi*

Korea Advanced Institute of Science and Technology, Daejeon, South Korea
{leesol4553, hyunjikim, sr5043, handonghee, munyi}@kaist.ac.kr

Abstract. Multimodal brain tumor segmentation typically leverages multiple MRI modalities, yet incomplete modality acquisition is common in clinical practice due to protocol heterogeneity and scan failures. Although recent methods maintain segmentation accuracy under missing modality conditions, they frequently overlook prediction reliability, leading to miscalibrated confidence estimates that hinder clinical adoption. Existing calibration techniques are largely modality-agnostic or assume that prediction difficulty decreases monotonically as additional modalities become available. However, in brain tumor segmentation, prediction difficulty depends primarily on which modalities are absent rather than how many, leading to combination-specific and spatially heterogeneous calibration errors. To address this, we propose **M**issing **M**odality-**A**ware **L**ocal **T**emperature **S**caling (**MMA-LTS**), a post-hoc voxel-wise confidence calibration method. It estimates a spatially adaptive temperature field conditioned on a modality-availability learnable token and a voxel-wise difficulty score. Experiments on BraTS 2020 and FeTS 2024 show that MMA-LTS improves calibration while preserving the segmentation accuracy of state-of-the-art models across diverse missing-modality scenarios, thereby enhancing trustworthiness toward clinical deployment.

Keywords: Brain Tumor Segmentation · Multi-Modal · Missing Modality · Confidence Calibration · Post-hoc Calibration

1 Introduction

Multimodal brain tumor segmentation plays a critical role in diagnosis, treatment planning, and prognosis assessment. Accurate delineation requires integrating complementary information from multiple MRI modalities, such as T1, T2, T1ce, and FLAIR [4]. However, in real-world clinical workflows, acquiring all modalities is often infeasible due to variations in imaging protocols, scan failures, or patient conditions [25]. As a result, models must operate robustly under incomplete modality settings. Prior approaches addressed this challenge by synthesizing missing modalities [16,23] or training separate models for each possible modality combination [8,22]. More recent methods achieve stronger performance

* Corresponding author.

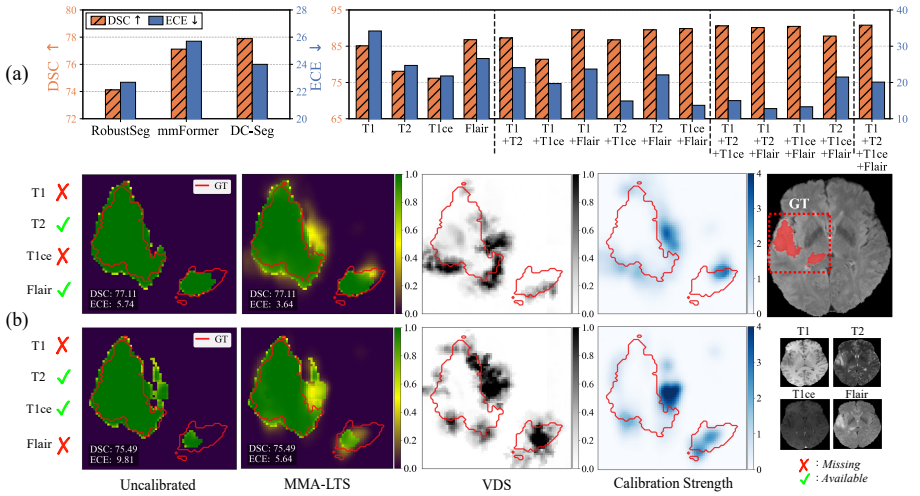

Fig. 1. Segmentation performance and calibration results. (a) Left: DSC and ECE of state-of-the-art models on BraTS 2020, averaged over all modality combinations. Right: DC-Seg results per modality combination for WT. (b) Two cases with the same modality count but differing in combination, showing DC-Seg confidence maps (uncalibrated and MMA-LTS), voxel-wise difficulty score (VDS), and calibration strength ($|T-1|$).

by embedding multimodal inputs into a shared latent space [2,10,26], enabling a unified architecture.

Despite recent advances, most studies prioritize segmentation accuracy while paying limited attention to prediction reliability under missing-modality conditions. Fig. 1(a) shows that the state-of-the-art DC-Seg [10] achieves higher Dice similarity coefficient (DSC) yet exhibits larger expected calibration error (ECE) than RobustSeg [2]. This suggests that high segmentation performance under incomplete multimodal inputs can obscure poorly calibrated confidence estimates that do not reflect true accuracy. Reliable confidence is essential in medical image segmentation, as miscalibration can undermine clinical decision-making [1]. This issue is particularly pronounced in brain tumor segmentation, where heterogeneous subregions and boundaries require calibrated confidence to identify unreliable delineations and support expert intervention [5,15].

Most prior work on segmentation calibration relies on joint prediction-calib ration methods [1,12,18] or post-processing probability correction [5,24], which assume fixed and fully observed inputs and therefore do not address multimodal or missing-modality settings. Although calibration methods for multimodal classification under missing modalities have been proposed [11,14], they typically assume prediction difficulty decreases and confidence increases monotonically as more modalities become available. This assumption does not hold in brain tumor segmentation, where modalities contribute unequally: T1ce primarily highlights enhancing tumor (ET) regions, whereas T2 and FLAIR better capture whole tumor (WT) extent [17]. As illustrated in Fig. 1(b), prediction difficulty depends

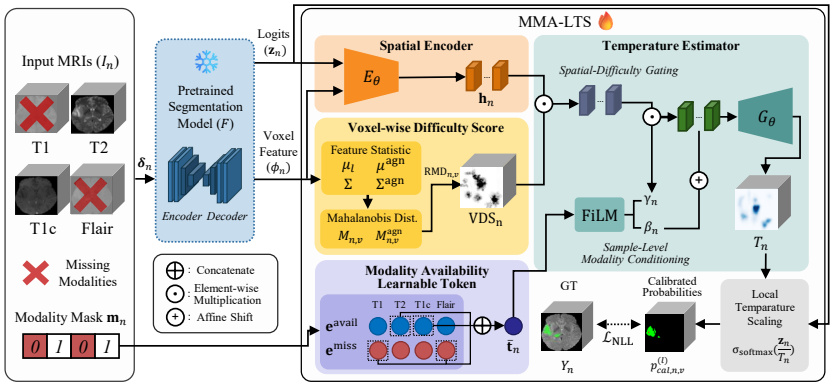

Fig. 2. The overall framework of MMA-LTS.

on which modalities are missing rather than how many, resulting in spatially heterogeneous calibration errors.

Motivated by this observation, we propose **MMA-LTS** (**M**issing **M**odality-**A**ware **L**ocal **T**emperature **S**caling), a plug-in post-hoc calibration method for multimodal segmentation with incomplete modalities. Unlike prior work, MMA-LTS performs voxel-wise calibration via modality-aware conditioning and spatial difficulty gating, correcting modality-combination-specific and spatially heterogeneous calibration errors without retraining. Experiments on BraTS 2020 [17] and FeTS 2024 [25] demonstrate improved calibration across missing-modality scenarios while preserving segmentation accuracy.

2 Method

Assume disjoint data splits $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and $\mathcal{D}_{\text{test}}$, each consisting of pairs (I_n, Y_n), where n is the sample index, $I_n \in \mathbb{R}^{|\mathcal{M}|\times D\times H\times W}$ is the multimodal input, $Y_n \in L^{D\times H\times W}$ is the ground-truth label, and $\mathcal{M}$ and L denote the sets of modalities and classes, respectively. A segmentation model F pretrained on $\mathcal{D}_{\text{train}}$ takes I_n and produces logits $\mathbf{z}_n \in \mathbb{R}^{|L|\times D\times H\times W}$. The proposed method aims to calibrate $\mathbf{z}_n$ of F on $\mathcal{D}_{\text{test}}$ under arbitrary modality availability. To this end, a calibration module is trained on $\mathcal{D}_{\text{val}}$, yielding spatially adaptive confidence estimates without retraining F. The overall framework is shown in Fig. 2.

2.1 Missing Modality-Aware Local Temperature Scaling

Spatial Encoder. F extracts modality-specific features $\{\mathbf{f}_{n,k}\}_{k\in\mathcal{M}}$ and produces a joint representation $\boldsymbol{\phi}_n \in \mathbb{R}^{C_f\times D\times H\times W}$ and $\mathbf{z}_n$ from I_n, where C_f is the channel dimension. These two signals are concatenated and passed through a 3D CNN module E_θ to obtain a spatial embedding $\mathbf{h}_n \in \mathbb{R}^{C_f\times D\times H\times W}$, which serves as the base representation for temperature estimation.

Voxel-wise Difficulty Score (VDS). To capture spatial difficulty induced by modality loss, we extend Mahalanobis-distance-based estimation [3] to 3D segmentation. We collect voxel feature-label pairs $\mathcal{V} = \{(\mathbf{g}_{n,v}, y_{n,v})\}$ from $\mathcal{D}_{\text{train}}$ to model the feature distribution of pretrained F, where v indexes a voxel location, $\mathbf{g}_{n,v} = \boldsymbol{\phi}_n(v) \in \mathbb{R}^{C_f}$ denotes the feature vector at voxel v, $y_{n,v} = Y_n(v) \in L$ the corresponding label, and $\mathcal{V}_l$ the subset of voxels belonging to each class $l \in L$. We estimate the class-conditional and class-agnostic statistics as:

$$\begin{aligned} \boldsymbol{\mu}_l &= \frac{1}{|\mathcal{V}_l|} \sum_{(n,v)\in\mathcal{V}_l} \mathbf{g}_{n,v}, \qquad \boldsymbol{\Sigma} = \frac{1}{|\mathcal{V}|} \sum_{l\in L} \sum_{(n,v)\in\mathcal{V}_l} (\mathbf{g}_{n,v} - \boldsymbol{\mu}_l)(\mathbf{g}_{n,v} - \boldsymbol{\mu}_l)^\top, \\ \boldsymbol{\mu}^{\text{agn}} &= \frac{1}{|\mathcal{V}|} \sum_{(n,v)\in\mathcal{V}} \mathbf{g}_{n,v}, \qquad \boldsymbol{\Sigma}^{\text{agn}} = \frac{1}{|\mathcal{V}|} \sum_{(n,v)\in\mathcal{V}} (\mathbf{g}_{n,v} - \boldsymbol{\mu}^{\text{agn}})(\mathbf{g}_{n,v} - \boldsymbol{\mu}^{\text{agn}})^\top. \end{aligned} \tag{1}$$

The class-conditional and agnostic Mahalanobis distances are then defined as:

$$\begin{aligned} M_{n,v} &= (\mathbf{g}_{n,v} - \boldsymbol{\mu}_{y_{n,v}})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{g}_{n,v} - \boldsymbol{\mu}_{y_{n,v}}), \\ M^{\text{agn}}_{n,v} &= (\mathbf{g}_{n,v} - \boldsymbol{\mu}_{\text{agn}})^\top (\boldsymbol{\Sigma}^{\text{agn}})^{-1} (\mathbf{g}_{n,v} - \boldsymbol{\mu}^{\text{agn}}). \end{aligned} \tag{2}$$

$M_{n,v}$ measures deviation from the class distribution and $M^{\text{agn}}_{n,v}$ from the global distribution, where $y_{n,v}$ is obtained from ground-truth labels during training and replaced by the predicted label $\hat{y}_{n,v} = \arg\max \mathbf{z}_{n,v}$ at inference time. We adopt $\text{RMD}_{n,v} = M_{n,v} - M^{\text{agn}}_{n,v}$ as a voxel-wise difficulty proxy, where larger values indicate voxels that are atypical within their class yet unremarkable globally, reflecting low class discriminability and thus indicating harder regions. After normalization and reshaping, we obtain $\text{VDS}_n \in [0,1]^{D\times H\times W}$.

Modality Availability Learnable Token. We encode modality availability as a sample-level conditioning signal for calibration. Let $\mathcal{S}(I_n) \subseteq \mathcal{M}$ denote the set of available modalities and define the binary mask $\mathbf{m}_{n,k} = \mathbb{I}[k \in \mathcal{S}(I_n)]$, where 1 and 0 indicate available and missing, respectively. For each $k \in \mathcal{M}$, learnable tokens $\mathbf{e}^{\text{avail}}_k, \mathbf{e}^{\text{miss}}_k \in \mathbb{R}^{d_e}$ represent the available and missing states, where d_e is the token dimension. The modality state token $\mathbf{t}_{n,k}$ is defined as

$$\mathbf{t}_{n,k} = \mathbf{m}_{n,k}\mathbf{e}^{\text{avail}}_k + (1 - \mathbf{m}_{n,k})\mathbf{e}^{\text{miss}}_k. \tag{3}$$

Concatenating all modality tokens yields the summary vector $\bar{\mathbf{t}}_n = \bigoplus_{k\in\mathcal{M}} \mathbf{t}_{n,k} \in \mathbb{R}^{|\mathcal{M}|\cdot d_e}$, which serves to globally modulate the calibration strength.

Temperature Estimator. The estimator learns a mapping that produces a voxel-wise temperature field using spatial features $\mathbf{h}_n$, voxel-wise difficulty score VDS_n, and modality availability token $\bar{\mathbf{t}}_n$. Spatial difficulty gating is first applied to selectively emphasize regions requiring stronger calibration. Sample-level modality conditioning is then performed via FiLM-based affine modulation [21] to adjust feature modulation according to the sample-wise modality configuration. The resulting temperature field $T_n \in \mathbb{R}^{D\times H\times W}$ is then given by

$$T_n = G_\theta\Big((\mathbf{h}_n \odot \text{VDS}_n) \odot \gamma_n + \beta_n\Big), \tag{4}$$

where $\odot$ denotes element-wise multiplication, $(\gamma_n, \beta_n) = f_{\text{FiLM}}(\bar{\mathbf{t}}_n)$ denote affine modulation parameters, and G_θ is a 3D CNN predictor. Spatial-difficulty gating determines *where* stronger calibration is needed, while sample-level modality conditioning controls *how strongly* it is applied based on available modalities.

Local Temperature Scaling. Following prior work [5], the estimated temperature field T_n is applied to post-hoc rescale the logits $\mathbf{z}_n$ of F via temperature-scaled softmax, yielding calibrated voxel-wise class probabilities:

$$p^{(l)}_{\text{cal},n,v} = \sigma_{\text{softmax}}\left(\frac{\mathbf{z}_{n,v}}{T_{n,v}}\right)^{(l)}, \quad l \in L. \tag{5}$$

where $T_{n,v}$ and $\mathbf{z}_{n,v}$ denote the temperature and logit at voxel v, respectively. Since $T_{n,v} \in \mathbb{R}^+$, class ordering is preserved, with values greater than 1 reducing overconfidence and values less than 1 mitigating underconfidence.

Training Objective. ModDrop [20] is applied as $\tilde{\mathbf{f}}_{n,k} = \delta_{n,k}\mathbf{f}_{n,k}$, where $\delta_{n,k}$ is a Bernoulli indicator randomly set to 0 during training. MMA-LTS is optimized via the negative log-likelihood (NLL) loss over the voxel grid Ω, which counteracts the miscalibration of F by recovering the true probability distribution:

$$\mathcal{L}_{\text{NLL}} = -\frac{1}{|\Omega|}\sum_{v\in\Omega} \log p^{(y_{n,v})}_{\text{cal},n,v} \tag{6}$$

3 Experiments

3.1 Implementation Details

Dataset. Experiments are conducted on BraTS 2020 [17] and FeTS 2024 [25], each providing 3D multi-contrast MRI scans with four modalities (FLAIR, T1ce, T1, T2) and three tumor subregions: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). BraTS 2020 contains 369 cases split as (219/50/100) for ($\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{val}}/\mathcal{D}_{\text{test}}$) [10], and FeTS 2024 contains 1,251 cases split as (1000/125/126) [27]. All volumes are skull-stripped, co-registered to a common anatomical template, resampled to $1mm^3$, and z-score normalized.

Segmentation Backbones and Calibration Methods. We evaluate MMA-LTS on state-of-the-art brain tumor segmentation models for missing-modality robustness: RobustSeg [2], mmFormer [26], and DC-Seg [10]. For calibration, we compare against joint prediction-calibration approaches such as mL1-ACE [1] and SDC [12], as well as post-hoc probability correction methods, including Temperature Scaling (TS) [7], Dirichlet Scaling (DirS) [9], Local Temperature Scaling (LTS) [5], Selective Scaling (SelS) [24], and MC Dropout (MC) [6]. All calibration methods are applied uniformly across backbones.

Training Settings. All backbones are trained on $\mathcal{D}_{\text{train}}$ with batch size 2 for 500/250 epochs on BraTS 2020/FeTS 2024, augmented with random 80^3 crops, $\pm 10^\circ$ rotation, intensity shifts, and flipping. Post-hoc calibration methods and MMA-LTS are trained for 40 epochs using AdamW [13] (batch size 1, learning rate 1×10^{-4}) on $\mathcal{D}_{\text{val}}$, while mL1-ACE and SDC are trained with backbones.

Table 1. DSC and ECE comparison of DC-Seg on BraTS 2020. • and ○ denote available and missing modalities. Temperature scaling-based methods (TS, DirS, LTS, SelS, MMA-LTS) preserve the DSC of the uncalibrated model; only ECE is reported.

Type	FLAIR	○	○	○	•	○	○	•	○	•	•	•	•	•	○	•	•	Avg.
	T1	○	○	•	○	○	•	•	•	○	○	•	•	○	•	•	•	
	T1ce	○	•	○	○	•	•	○	○	○	•	•	○	•	•	•		
	T2	•	○	○	○	•	○	○	•	•	○	○	•	•	•	•		
WT	Uncalibrated	24.7	21.8	34.2	26.6	14.9	19.7	23.7	24.1	22.1	13.7	13.3	21.5	12.8	15.0	13.0	20.1
	⌞ DSC	85.1	78.1	76.2	86.8	87.3	81.4	89.5	86.7	89.5	89.9	90.6	90.1	90.5	87.8	90.8	86.7
	mL1-ACE [1]	27.0	25.4	39.4	26.4	16.4	23.5	23.7	25.9	22.1	14.0	13.8	21.2	13.0	16.4	13.0	21.4
	⌞ DSC	83.7	76.9	72.4	86.1	86.6	79.4	88.2	85.1	88.5	89.2	89.9	89.5	90.0	86.8	90.5	85.5
	SDC [12]	22.4	21.8	33.4	23.5	13.7	19.2	20.4	21.8	19.0	12.7	12.2	18.4	11.7	13.8	11.6	18.4
	⌞ DSC	85.6	78.3	75.6	86.8	87.7	81.5	89.5	87.1	89.8	90.1	90.8	90.4	90.7	88.1	91.0	86.9
	MC [6]	24.8	22.0	34.2	26.7	14.9	19.7	23.6	24.3	22.2	13.6	13.4	21.5	12.9	15.3	13.0	20.1
	⌞ DSC	85.0	77.9	76.0	86.7	87.3	81.2	89.4	86.4	89.3	89.8	90.6	90.2	90.4	87.7	90.9	86.6
	TS [7]	23.9	21.1	33.2	25.8	14.3	19.0	22.8	23.2	21.4	13.1	12.7	20.7	12.3	14.4	12.5	19.4
	DirS [9]	24.9	22.9	34.7	26.6	15.6	20.6	23.4	24.3	22.1	14.1	13.5	21.4	13.3	15.6	13.2	20.4
	LTS [5]	14.2	16.2	22.9	16.2	10.3	14.2	14.0	13.9	13.1	10.2	9.7	12.4	9.5	10.4	9.4	13.1
	SelS [24]	20.4	18.2	29.3	22.9	11.3	15.2	19.9	19.8	18.7	10.7	10.4	18.1	9.9	11.2	10.1	16.3
	MMA-LTS*	**12.6**	**12.7**	**15.1**	**13.0**	**8.1**	**10.2**	**11.8**	**11.2**	**11.4**	**8.2**	**7.2**	**10.1**	**7.0**	**7.3**	**6.8**	**10.2**
TC	Uncalibrated	37.6	19.1	43.2	43.4	16.5	17.9	38.9	35.6	35.5	16.4	17.4	35.4	16.1	16.9	17.0	27.1
	⌞ DSC	73.2	84.5	64.6	70.7	87.6	86.8	73.9	73.9	75.3	87.4	87.4	75.5	87.5	87.2	87.3	80.2
	mL1-ACE	33.6	17.6	40.7	38.8	14.0	16.6	34.3	32.2	31.1	13.4	14.2	30.7	12.9	14.3	13.8	23.9
	⌞ DSC	72.5	84.0	63.2	70.0	87.4	84.9	72.8	73.0	74.4	87.5	87.4	74.9	87.6	86.7	87.1	79.6
	SDC	31.8	16.7	37.6	36.3	14.0	15.0	32.2	30.1	29.1	14.0	14.7	29.1	13.5	14.5	14.4	22.9
	⌞ DSC	73.3	84.7	64.3	70.5	87.6	86.0	74.1	74.0	75.8	87.1	87.1	76.0	87.2	86.9	87.0	80.1
	MC	37.6	19.4	42.8	43.7	16.6	17.8	38.6	35.8	35.6	16.4	17.5	35.3	16.0	17.0	17.0	27.1
	⌞ DSC	73.0	84.8	54.8	70.7	87.6	86.9	74.3	73.7	75.4	87.5	87.5	75.5	87.5	87.4	87.4	80.3
	TS	36.5	18.4	41.9	42.2	15.9	17.2	37.8	34.6	34.4	15.8	16.8	34.4	15.5	16.3	16.4	26.3
	DirS	34.9	17.6	40.5	40.7	15.2	16.3	36.1	33.0	32.9	15.1	15.8	32.7	14.8	15.5	15.5	25.1
	LTS	22.2	12.6	27.5	27.7	10.6	11.1	24.1	21.2	21.2	10.2	10.5	21.6	10.0	10.9	10.7	16.8
	SelS	32.2	15.7	37.2	39.8	14.3	15.5	35.1	32.0	32.2	14.0	15.0	32.0	13.9	14.7	13.5	24.2
	MMA-LTS*	**18.2**	**10.9**	**20.4**	**20.1**	**10.2**	**10.1**	**20.9**	**18.9**	**17.6**	**9.7**	**10.3**	**19.3**	**9.7**	**10.6**	**10.6**	**14.5**
ET	Uncalibrated	36.5	12.8	46.8	46.9	9.9	10.9	43.4	37.2	37.5	11.2	10.5	38.0	10.4	10.1	10.4	24.8
	⌞ DSC	51.8	80.0	40.9	42.4	83.2	83.4	47.0	51.6	52.5	82.7	83.5	52.6	83.2	83.6	83.6	66.8
	mL1-ACE	31.8	11.1	45.3	43.0	7.8	10.5	39.6	34.4	32.9	8.5	7.8	33.6	7.9	8.0	7.9	22.0
	⌞ DSC	51.4	79.4	38.6	41.8	83.2	81.2	45.8	50.3	51.9	82.7	83.5	52.0	83.1	83.3	83.3	66.1
	SDC	11.8	31.5	42.7	38.0	8.9	9.7	36.2	32.6	31.4	9.5	8.9	32.1	8.9	8.9	9.0	21.3
	⌞ DSC	51.3	80.0	39.1	44.0	83.3	83.3	47.7	51.0	54.7	82.8	83.5	52.7	83.2	83.5	83.4	83.5
	MC	36.4	13.1	46.6	47.6	10.1	10.5	43.0	37.4	37.6	11.1	10.6	37.7	10.3	10.0	10.3	24.8
	⌞ DSC	51.4	80.8	40.5	41.4	83.4	83.8	46.7	51.1	51.8	83.2	83.6	52.3	83.7	84.0	83.7	66.8
	TS	35.3	12.3	45.5	45.7	9.5	10.4	42.2	36.1	36.4	10.7	10.0	36.9	10.0	9.6	9.9	24.0
	DirS	33.3	12.0	43.5	43.5	9.1	10.0	39.7	33.9	34.2	10.3	9.4	34.6	9.5	9.1	9.3	22.8
	LTS	21.5	9.4	30.6	31.7	7.5	7.6	29.3	**22.9**	23.4	8.1	7.7	24.4	7.7	7.6	7.7	16.5
	SelS	30.9	11.2	40.4	42.7	8.6	9.1	39.2	32.9	33.5	9.8	9.2	33.9	9.3	8.7	9.1	22.1
	MMA-LTS*	**21.1**	**8.8**	**25.1**	**27.0**	**7.1**	**7.0**	**28.0**	**22.9**	**21.9**	**7.8**	**6.9**	**23.4**	**6.8**	**6.6**	**6.4**	**15.1**

MC uses $p = 0.3$ with 5 forward passes. For MMA-LTS, we set $C_f = 16$ and $d_e = 256$. All experiments are conducted on an NVIDIA® RTX A6000 GPU.

Evaluation Metrics. We evaluate segmentation with DSC [28] and calibration with ECE [19] following recent works [1,12], where ECE is computed over foreground voxels using 10 bins and averaged across test volumes.

3.2 Experiment Results

Comparative Experiments. Tables 1 and 2 present results on BraTS 2020 and FeTS 2024. MMA-LTS achieves the lowest ECE across most tumor regions and 15 modality combinations while preserving DSC. On BraTS 2020, MMA-LTS reduces average WT/TC/ET ECE to 10.2/14.5/15.1, outperforming the second-best LTS (13.1/16.8/16.5). On FeTS 2024, MMA-LTS similarly achieves the best ECE across tumor regions (9.5/12.0/11.6), demonstrating effective correction of modality-combination-induced and spatially heterogeneous confidence distortions. Table 3 further confirms consistent ECE reduction across RobustSeg and mmFormer backbones, validating its generalized calibration effectiveness.

Ablation Study. Fig. 3 presents ablation results of MMA-LTS on DC-Seg

Table 2. DSC and ECE comparison of DC-Seg on FeTS 2024.

Type	FLAIR	○	○	○	●	○	○	●	○	●	●	●	●	●	○	●	Avg.
	T1	○	○	●	○	○	●	●	●	○	○	●	●	○	●	●	
	T1ce	○	●	○	○	●	●	○	○	○	●	●	○	●	●	●	
	T2	●	○	○	○	●	○	○	●	●	○	○	●	●	●	●	
WT	Uncalibrated	20.9	17.0	25.3	17.9	12.1	14.7	15.7	19.3	15.3	8.7	8.5	14.7	8.5	12.0	14.6	15.0
	∟ DSC	87.7	79.3	80.8	90.3	89.0	83.4	91.8	88.5	91.9	91.6	92.3	92.2	92.5	89.2	88.9	88.6
	mL1-ACE [1]	20.4	20.7	27.2	15.9	12.7	17.4	14.0	18.5	13.8	8.1	8.1	13.2	7.9	12.5	8.0	14.6
	∟ DSC	86.5	78.8	78.4	90.5	88.1	82.3	91.7	87.3	91.9	92.1	92.4	92.1	92.6	88.4	92.6	88.4
	SDC	18.2	17.3	22.6	15.2	10.9	13.6	13.1	16.3	13.2	7.6	7.5	12.3	7.3	10.7	7.3	10.7
	∟ DSC	87.8	80.1	80.5	90.8	89.2	83.7	92.1	88.7	92.3	92.1	92.6	92.4	92.8	89.3	92.8	89.1
	MC [6]	20.8	16.9	25.3	17.9	12.0	14.7	15.8	19.3	15.3	8.7	8.5	14.7	8.5	12.0	14.6	15.0
	∟ DSC	87.7	79.3	80.6	90.3	89.0	83.4	92.0	88.4	92.1	91.8	92.4	92.4	92.6	89.1	92.7	88.9
	TS [7]	18.4	15.3	22.2	15.6	10.8	13.1	13.5	16.9	13.2	7.6	7.4	12.7	7.4	10.7	12.8	13.2
	DirS [9]	19.9	18.0	25.5	16.3	12.7	15.6	14.3	18.5	14.0	8.5	8.5	13.5	8.5	12.6	14.3	14.7
	LTS [5]	12.2	14.0	15.9	12.1	9.8	11.7	10.0	11.8	9.9	8.8	8.6	9.7	8.8	10.0	8.7	10.9
	SelS [24]	15.1	13.6	17.8	13.0	9.1	12.0	11.2	13.7	11.1	6.6	6.9	10.7	7.0	9.0	10.9	11.2
	MMA-LTS*	**11.8**	**13.2**	**15.7**	**11.9**	**8.5**	**10.3**	**9.9**	**10.3**	**8.7**	**6.9**	**6.5**	**8.1**	**6.3**	**8.4**	**6.1**	**9.5**
TC	Uncalibrated	30.2	14.5	32.4	31.8	10.9	13.0	26.3	25.9	25.4	11.6	11.5	23.7	10.6	10.8	19.3	19.9
	∟ DSC	75.8	88.9	73.8	76.9	91.6	90.5	80.4	80.4	81.0	91.3	91.8	82.3	92.0	92.2	85.4	84.3
	mL1-ACE	26.5	13.9	29.3	27.2	9.9	12.1	22.0	22.9	23.2	10.7	10.3	21.3	9.9	10.4	9.7	17.3
	∟ DSC	74.3	88.0	71.4	75.5	90.9	89.4	78.7	77.7	78.7	90.6	91.0	80.0	91.1	90.6	91.3	84.9
	SDC	23.4	12.5	28.7	27.6	9.7	11.1	22.9	21.9	21.9	10.0	10.2	20.3	9.5	9.7	9.5	16.7
	∟ DSC	76.2	89.2	73.1	77.1	91.6	90.6	80.2	80.1	80.9	91.7	91.9	82.3	91.9	92.1	92.2	85.4
	MC	30.1	14.4	32.7	31.7	10.8	12.9	26.3	25.7	25.2	11.7	11.5	23.6	10.6	10.9	19.2	19.8
	∟ DSC	75.9	88.9	73.8	76.9	91.5	90.4	80.3	80.5	81.2	91.4	91.9	82.7	92.0	92.1	92.4	85.5
	TS	26.9	13.3	28.6	28.4	9.8	11.9	23.1	22.7	22.4	10.5	10.4	20.8	9.5	9.7	17.2	17.7
	DirS	26.7	13.2	28.5	28.2	9.9	11.7	23.0	22.7	22.1	10.3	10.3	20.7	9.4	9.6	17.0	17.6
	LTS	18.2	14.4	19.9	17.7	11.7	13.2	15.4	14.5	14.7	11.9	11.7	14.2	11.3	11.4	11.4	14.1
	SelS	23.7	12.5	23.3	24.3	9.2	11.6	20.3	19.5	20.7	9.8	9.8	18.9	9.2	9.2	15.8	15.8
	MMA-LTS*	**17.7**	**11.3**	**19.5**	**17.6**	**7.8**	**9.4**	**15.2**	**14.1**	**14.3**	**8.1**	**8.0**	**13.6**	**7.5**	**7.6**	**7.8**	**12.0**
ET	Uncalibrated	27.4	10.2	31.4	27.8	7.8	8.9	24.2	23.6	22.6	7.7	7.5	21.5	7.6	7.8	16.2	16.8
	∟ DSC	58.9	85.5	54.6	59.6	87.6	87.6	63.5	63.7	65.3	87.7	88.6	66.8	88.0	88.3	75.6	74.5
	mL1-ACE	23.8	10.7	28.7	24.0	7.5	8.9	20.4	21.0	21.1	7.4	6.9	19.5	6.9	7.7	6.5	14.7
	∟ DSC	58.0	84.6	52.6	58.1	87.1	86.3	62.3	61.4	63.8	87.4	88.1	65.5	87.8	86.9	88.2	74.5
	SDC	23.1	9.0	27.8	23.8	6.7	7.7	21.0	19.8	19.5	6.9	6.8	18.5	6.7	6.7	6.7	14.1
	∟ DSC	59.1	85.5	54.1	59.3	87.7	87.5	63.4	63.4	65.2	87.8	88.6	66.9	88.1	88.4	88.5	75.6
	MC	27.5	10.1	31.7	27.7	7.7	8.8	24.3	23.5	22.5	7.7	7.5	21.3	7.6	7.8	16.2	16.8
	∟ DSC	58.9	85.5	54.0	59.3	87.5	87.5	63.0	63.6	65.3	87.7	88.6	66.9	88.0	88.2	88.5	75.5
	TS	24.2	9.2	27.6	24.6	6.9	8.0	21.2	20.5	19.8	6.8	6.6	18.8	6.6	6.8	14.3	14.8
	DirS	24.3	9.3	27.8	24.7	7.1	8.0	21.4	20.9	19.8	6.8	6.6	18.9	6.7	6.9	14.4	14.9
	LTS	20.3	11.4	21.8	**18.6**	9.9	9.6	17.9	**16.3**	16.5	8.8	8.7	17.0	8.6	8.6	8.3	13.5
	SelS	24.4	8.8	24.8	24.0	6.5	7.7	20.0	20.2	19.0	6.5	6.4	18.7	6.2	6.5	13.8	13.1
	MMA-LTS*	**20.1**	**8.6**	**21.7**	19.1	**5.9**	**6.7**	**17.8**	16.8	**15.6**	**5.3**	**5.1**	**15.3**	**5.0**	**5.5**	**5.1**	**11.6**

Table 3. DSC and ECE comparison of RobustSeg and mmFormer on BraTS 2020 and FeTS 2024. Values are reported as ECE/DSC and averaged over modality combinations.

Dataset	Model	Type	Uncal	mL1-ACE	SDC	MC	TS	DirS	LTS	SelS	MMA-LTS*
BraTS 2020	RobustSeg	WT	20.7/84.1	19.7/80.1	20.1/83.5	20.7/84.3	19.5	20.5	15.6	15.9	**14.9**
		TC	26.2/75.3	24.0/71.3	23.7/74.8	25.5/75.4	24.9	24.0	18.9	18.6	**16.6**
		ET	21.2/63.0	19.9/59.8	19.3/62.8	20.8/63.1	20.2	20.1	21.0	20.0	**15.8**
	mmFormer	WT	22.1/86.3	20.3/86.1	20.4/86.6	22.2/86.1	21.5	22.1	17.0	18.4	**11.6**
		TC	29.1/79.7	25.3/78.7	26.3/79.7	29.2/79.6	28.4	28.4	19.4	24.9	**17.4**
		ET	26.5/65.4	22.6/64.2	23.9/65.3	26.6/65.2	25.8	26.3	17.8	22.1	**16.5**
FeTS 2024	RobustSeg	WT	15.9/86.7	17.6/85.3	15.0/86.9	16.5/86.9	13.9	15.2	11.4	10.9	**9.5**
		TC	20.8/81.6	19.4/79.5	17.6/81.7	20.9/81.7	18.5	18.5	14.3	14.9	**12.9**
		ET	17.0/71.7	16.6/69.8	14.7/71.2	17.1/71.8	14.9	15.2	12.8	13.2	**12.0**
	mmFormer	WT	16.6/86.8	14.5/87.4	11.9/86.7	16.8/86.5	15.3	16.7	12.8	16.3	**11.8**
		TC	22.9/83.3	15.6/82.8	15.5/83.6	23.0/83.1	21.4	22.5	16.9	22.3	**13.3**
		ET	19.5/72.7	13.8/72.3	14.0/72.6	19.6/72.6	28.0	21.2	17.2	19.0	**13.2**

with BraTS 2020. Removing VDS or replacing it with sample-averaged difficulty (SDS) consistently increases ECE across all tumor regions, highlighting the importance of spatial difficulty estimation. Removing the modality availability token $\bar{t}_n$ or substituting it with a count-based token also degrades ECE, as it is limited in capturing modality combination-specific calibration errors. These results demonstrate that each component contributes to calibration improvement across all tumor regions under missing modalities.

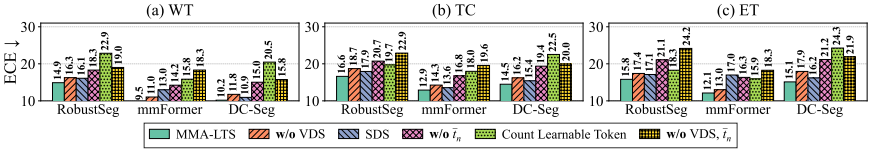

Fig. 3. Ablation study of MMA-LTS. ECE are averaged over modality combinations.

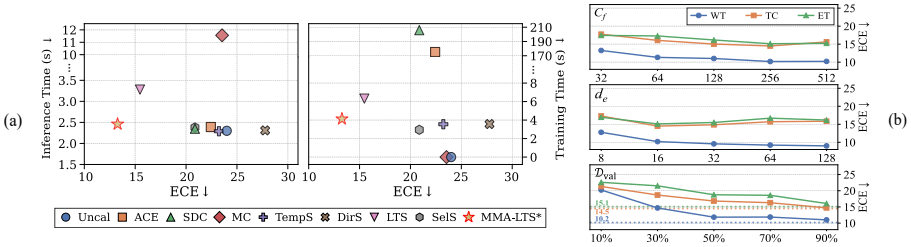

Fig. 4. Time efficiency and sensitivity analysis (DC-Seg, BraTS 2020). (a) ECE versus inference and training time. (b) Sensitivity to C_f, d_e, and $\mathcal{D}_{\text{val}}$ size (dashed: 100%).

Qualitative Analysis. Fig. 1(b) presents two cases with identical modality counts but different modality combinations (T2+FLAIR and T2+T1ce). Although their DSC values are comparable, the uncalibrated model exhibits different confidence patterns, with pronounced overconfident errors near tumor boundaries, assigning high tumor probability to non-tumor regions (false positives) and confidently mispredicting true tumors as background (false negatives), particularly in the T2+T1ce case. The VDS maps consistently highlight these ambiguous regions, while the calibration strength maps indicate where stronger correction is required for different modality combinations. MMA-LTS suppresses overconfident predictions and corrects spatially heterogeneous confidence distortions, yielding more reliable confidence estimates. Nonetheless, residual overconfidence near tumor boundaries remains, leaving room for improvement.

Time Efficiency and Sensitivity Analysis. As a lightweight post-hoc module with a frozen backbone, MMA-LTS achieves the lowest ECE with inference time comparable to temperature scaling-based methods and much lower training cost than retraining approaches (mL1-ACE, SDC) (Fig. 4(a)). MC requires multiple forward passes and thus higher inference cost. Sensitivity analysis (Fig. 4(b)) confirms robustness to hyperparameters and $\mathcal{D}_{\text{val}}$ size, showing stable ECE for $C_f \geq 128$ and $d_e \geq 16$, with performance saturating beyond 50% of $\mathcal{D}_{\text{val}}$.

4 Conclusion

We present MMA-LTS, a plug-in post-hoc calibration method that corrects modality-combination-specific and spatially heterogeneous miscalibration in brain

tumor segmentation under missing modalities. By conditioning on modality availability and voxel-wise difficulty, MMA-LTS enables spatially adaptive calibration without retraining the segmentation model or compromising accuracy. Experiments on BraTS 2020 and FeTS 2024 demonstrate consistent ECE reductions across multiple backbones and diverse missing-modality scenarios, establishing MMA-LTS as a practical step toward trustworthy clinical deployment. Future work includes extending validation to other anatomical domains and imaging modalities, incorporating reliability metrics beyond ECE with formal statistical testing, and further clinical validation in real-world settings.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) under grant number RS-2022-NR068758. We are also deeply grateful for the generous support provided by the Seegene Medical Foundation in South Korea.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Barfoot, T., Garcia Peraza Herrera, L.C., Glocker, B., Vercauteren, T.: Average calibration error: A differentiable loss for improved reliability in image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 139–149. Springer (2024)
2. Chen, C., Dou, Q., Jin, Y., Chen, H., Qin, J., Heng, P.A.: Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In: International conference on medical image computing and computer-assisted intervention. pp. 447–456. Springer (2019)
3. Cui, P., Zhang, D., Deng, Z., Dong, Y., Zhu, J.: Learning sample difficulty from pretrained models for reliable prediction. Advances in Neural Information Processing Systems **36**, 25390–25408 (2023)
4. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3975–3984 (2021)
5. Ding, Z., Han, X., Liu, P., Niethammer, M.: Local temperature scaling for probability calibration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6889–6899 (2021)
6. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)
7. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
8. Hu, M., Maillard, M., Zhang, Y., Ciceri, T., La Barbera, G., Bloch, I., Gori, P.: Knowledge distillation from multi-modal to mono-modal segmentation networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 772–781. Springer (2020)

9. Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., Flach, P.: Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. Advances in neural information processing systems **32** (2019)
10. Li, H., Li, Z., Mao, Y., Ding, Z., Huang, Z.: Dc-seg: Disentangled contrastive learning for brain tumor segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 138–148. Springer (2025)
11. Li, S., Chen, C., Han, J.: Simmlm: A simple framework for multi-modal learning with missing modality. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24068–24077 (2025)
12. Liang, W., Zhang, W.E., Yue, L., Xu, M., Maennel, O., Chen, W.: Calibrating on medical segmentation model through signed distance. In: Proceedings of the 34th ACM International Conference on Information and Knowledge Management. pp. 1778–1787 (2025)
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
14. Ma, H., Zhang, Q., Zhang, C., Wu, B., Fu, H., Zhou, J.T., Hu, Q.: Calibrating multimodal learning. In: International Conference on Machine Learning. pp. 23429–23450. PMLR (2023)
15. Mehrtash, A., Wells, W.M., Tempany, C.M., Abolmaesumi, P., Kapur, T.: Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. IEEE transactions on medical imaging **39**(12), 3868–3878 (2020)
16. Meng, X., Sun, K., Xu, J., He, X., Shen, D.: Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. IEEE Transactions on Medical Imaging **43**(7), 2587–2598 (2024)
17. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
18. Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., Dokania, P.: Calibrating deep neural networks using focal loss. Advances in neural information processing systems **33**, 15288–15299 (2020)
19. Naeini, M.P., Cooper, G., Hauskrecht, M.: Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI conference on artificial intelligence. vol. 29 (2015)
20. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2015)
21. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
22. Qiu, Y., Chen, D., Yao, H., Xu, Y., Wang, Z.: Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21317–21326 (2023)
23. Shen, L., Zhu, W., Wang, X., Xing, L., Pauly, J.M., Turkbey, B., Harmon, S.A., Sanford, T.H., Mehralivand, S., Choyke, P.L., et al.: Multi-domain image completion for random missing input data. IEEE transactions on medical imaging **40**(4), 1113–1122 (2020)

24. Wang, D., Gong, B., Wang, L.: On calibrating semantic segmentation models: Analyses and an algorithm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23652–23662 (2023)
25. Zhang, C., Chen, Y., Fan, Z., Huang, Y., Weng, W., Ge, R., Zeng, D., Wang, C.: Tc-diffrecon: Texture coordination mri reconstruction method based on diffusion model and modified mf-unet method. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI). pp. 1–5. IEEE (2024)
26. Zhang, Y., He, N., Yang, J., Li, Y., Wei, D., Huang, Y., Zhang, Y., He, Z., Zheng, Y.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 107–117. Springer (2022)
27. Zhu, S., Chen, Y., Chen, W., Wang, Y., Liu, C., Jiang, S., Qin, F., Wang, C.: Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 95–106. Springer (2025)
28. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in mr images: method and validation. IEEE transactions on medical imaging **13**(4), 716–724 (1994)