MODELING UNCERTAINTY FROM MISSING MODALITIES FOR BRAIN TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Brain tumor segmentation from multimodal MRIs is challenging when one or more modalities are missing. Many existing methods focus on multimodal representation learning or knowledge distillation but pay limited attention to how missing modalities affect predictions and the uncertainty associated with them. In this work, we study uncertainty from missing modalities under partial observation. From a probabilistic perspective, we characterize uncertainty from missing modalities by the total conditional variance of the posterior under full observation and show that this quantity equals the expected squared change between the class posteriors under full and partial observations. Building on this insight, we adopt Dirichlet modeling for predictions under missing modalities and learn the Dirichlet variance by matching it to the discrepancy between predictions under full and partial observations. This allows the estimated uncertainty to reflect the effect of missing modalities on the current prediction. Extensive experiments on BraTS2020 and BraTS2024 demonstrate that our method achieves strong performance across diverse missing-modality settings. Uncertainty analyses show closer agreement between the estimated uncertainty and the prediction change caused by missing modalities.

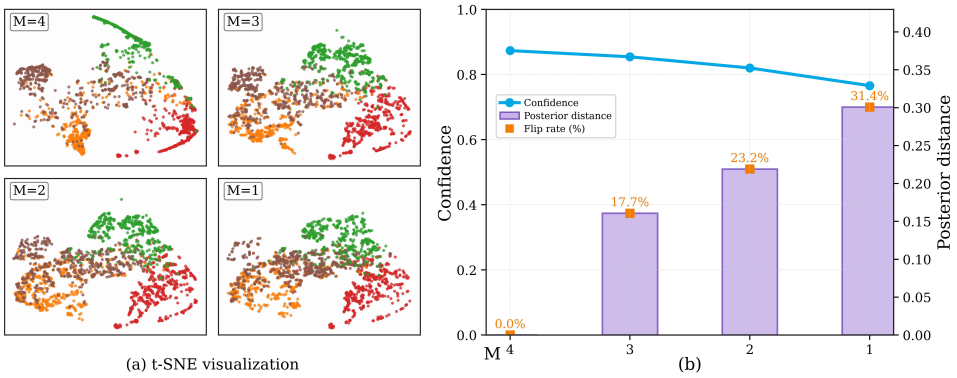
1 INTRODUCTION

Brain tumors pose a serious threat to human health. Magnetic Resonance Imaging (MRI) provides complementary tissue contrasts and pathological characteristics across modalities. Combining multiple MRI modalities enables more comprehensive tumor characterization and more reliable segmentation, supporting clinical diagnosis and treatment (Menze et al., 2015). In clinical practice, one or more MRI modalities may be missing due to variations in acquisition protocols, scanner availability or acquisition costs (Zhou et al., 2023). Maintaining segmentation performance while making full use of the available modalities has become an important problem in multimodal medical image segmentation.

Recent studies have improved segmentation with missing modalities through modality completion (Hamghalam et al., 2021; Yang et al., 2023), multimodal representation learning (Ding et al., 2021; Zhang et al., 2022) and knowledge distillation (Choi et al., 2023; Zhu et al., 2026). Although these methods use different learning strategies, they are designed to maintain segmentation performance when modalities are missing. As illustrated in Fig. 1, the posterior distance between predictions under full and partial observations increases as fewer modalities are available. The average prediction confidence remains around 0.76 even when only one modality is available. The corresponding posterior distance reaches approximately 0.30. This contrast suggests that prediction confidence alone does not adequately reflect the effect of missing modalities on the current prediction.

To address this issue, we examine how missing modalities affect model predictions and analyze prediction uncertainty when multimodal inputs are partially observed. As demonstrated in Sec. 3, uncertainty from missing modalities can be characterized by the expected squared discrepancy between the posterior distributions under full and partial observations. This analysis provides a probabilistic basis for using the discrepancy between predictions under full and partial observations as supervision during training, allowing the model to learn uncertainty that corresponds to the prediction changes caused by missing modalities.

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



(a) t-SNE visualization (b)

Figure 1: Effect of available modalities on posterior predictions and confidence. M denotes the number of available modalities. (a) t-SNE visualization of posterior probability vectors for $M = 4, 3, 2, 1$. Each panel aggregates all modality combinations with the corresponding M and each point represents a sampled voxel. Colors denote the four segmentation classes. The posterior distributions become less separated as M decreases. (b) Confidence, posterior distance and flip rate for different values of M. Confidence is the maximum predicted class probability under partial observation. Prediction distance is the squared Euclidean distance between the probability vectors under full and partial observations. Flip rate is the proportion of voxels whose predicted class differs between full and partial observations. Values are averaged over modality combinations with the same M.

Building on this analysis, we develop a joint learning framework using full and partial observations during training. Knowledge distillation uses predictions under full observation to guide predictions under partial observation. We also use the Brier score (Silva Filho et al., 2023) to encourage the predicted class probabilities under both observations to better approximate their corresponding posteriors. For uncertainty modeling, we introduce a Dirichlet distribution (Sensoy et al., 2018) for predictions under missing modalities and estimate a learnable concentration parameter from the features of the network. The resulting Dirichlet variance is learned by matching it to the prediction discrepancy between full and partial observations. At inference, both segmentation and uncertainty estimation rely only on the available modalities.

We validate our method on BraTS2020 and BraTS2024 across different combinations of missing modalities. The proposed method achieves competitive segmentation performance. We further evaluate the learned uncertainty by examining its agreement with the prediction variation between full and partial observations. Spatial visualizations further illustrate the correspondence between the estimated uncertainty and prediction changes caused by missing modalities. Our main contributions are summarized as follows:

- We provide a probabilistic analysis of prediction uncertainty under partial observation and establish its relationship with the expected squared discrepancy between the posterior distributions under full and partial observations.
- We introduce Dirichlet modeling for uncertainty estimation and learn the Dirichlet variance by matching it to the prediction change between full and partial observations.
- We empirically validate the proposed method on two brain tumor segmentation benchmarks, demonstrating competitive segmentation performance and effective characterization of uncertainty from missing modalities.

2 Related Work

2.1 Brain Tumor Segmentation with Missing Modalities

Multimodal MRI provides complementary anatomical and pathological information for brain tumor segmentation, but missing modalities can substantially degrade segmentation performance in practice. Existing methods address this problem mainly through modality completion or multimodal representation learning (Zeng et al., 2024; Li et al., 2024; Jeong et al., 2022). Some methods recover

missing content by learning shared representations and representations specific to each modality (Li et al., 2025a; Liu et al., 2025a; Diao et al., 2025) while more recent approaches directly fuse the available modalities using attention or Transformer architectures to improve representation across different modality combinations (Wang & Hong, 2023; Ting & Liu, 2024; Qiu et al., 2024). Knowledge distillation provides another solution by transferring features or predictions from a model with all available modalities to a model operating with missing modalities, reducing the performance gap between different modality conditions (Zhu et al., 2025; Liu et al., 2025b; Zheng et al., 2026). Despite their different strategies for recovery, representation learning and knowledge transfer, these methods aim to compensate for the loss of useful content caused by missing modalities and preserve segmentation performance. In contrast, how missing modalities affect the current prediction and how this effect should be reflected by uncertainty remain less explored.

2.2 Uncertainty Estimation in Medical Image Segmentation

Uncertainty estimation has been widely studied for assessing the reliability of predictions in medical image segmentation (Huang et al., 2024). Existing approaches include Bayesian inference (Zhao et al., 2022), Monte Carlo dropout (Camarasa et al., 2020), model ensembles (Buddenkotte et al., 2023) and confidence calibration (Mehrtash et al., 2020). These approaches assess predictive reliability through posterior variability, variation across stochastic predictions, disagreement among models or the calibration of predicted probabilities. Evidential deep learning has recently been introduced to medical image segmentation. These methods typically use a Dirichlet distribution to model class probabilities and predictive uncertainty (Li et al., 2023; Zou et al., 2022). They have also been explored for confidence calibration, error detection and robustness analysis (Li et al., 2023; Zou et al., 2025; Li et al., 2025b). However, most existing approaches focus on general sources of uncertainty, such as prediction errors, class ambiguity, model calibration and distribution shift. The relationship between prediction changes caused by missing modalities and the learned uncertainty has received considerably less attention. We address this problem by analyzing prediction uncertainty under partial observation. We further develop a Dirichlet uncertainty learning objective that aligns the estimated uncertainty with the prediction changes caused by missing modalities.

3 Theoretical Motivation

Our goal is not only to maintain segmentation performance when modalities are missing, but also to characterize how missing modalities affect the current prediction. To this end, we analyze the relationship between the class posteriors under full and partial observations from a probabilistic perspective and define uncertainty from missing modalities accordingly. Based on this analysis, we further establish that the prediction change between full and partial observations provides an observable supervision signal for learning this uncertainty during training. Detailed proofs are provided in Appendix A.

3.1 Class Posteriors under Partial Observation

Let $X = (X^{(1)}, \ldots, X^{(K)})$ denote the full multimodal MRI observation, K denoting the number of modalities. In the brain tumor segmentation task considered in this work, $K = 4$, corresponding to T1, T1ce, T2 and FLAIR. Let $M = (M_1, \ldots, M_K) \in \{0, 1\}^K$ denote the modality availability pattern. The condition $M_k = 1$ indicates that the k-th modality is available. For a given M, the corresponding masked input is $X^M = T_M(X)$.

A partial observation contains the available modalities together with their availability pattern and is represented as $O = (X^M, M)$. In our implementation, each MRI modality occupies a fixed input channel. A missing modality is represented by setting the corresponding channel to zero. The modality availability pattern is therefore implicitly encoded in the input representation.

For an arbitrary voxel v, let $Y_v \in \{1, \ldots, C\}$ denote its class label. For simplicity, we omit the voxel index v in what follows. The class posterior vectors under full and partial observations are defined as

$$\boldsymbol{\eta}_F(X) = P(Y = \cdot \mid X), \qquad \boldsymbol{\eta}_M(O) = P(Y = \cdot \mid O).$$

We consider both randomly generated and predefined availability patterns. Conditioned on the full observation X, the availability pattern M provides no additional information about the class label Y. We therefore assume

$$Y \perp M \mid X. \tag{1}$$

This assumption is consistent with the random modality masking protocol used in our training procedure.

Proposition 1 (Conditional mean under partial observation). Under the conditional independence assumption in Eq. 1, the class posterior under partial observation equals the conditional mean of the class posterior under full observation given the partial observation:

$$\boldsymbol{\eta}_M(O) = \mathbb{E}[\boldsymbol{\eta}_F(X) \mid O]\,. \tag{2}$$

The proof of Proposition 1 is provided in Appendix A.1.

Eq. 2 establishes the probabilistic relationship between the class posteriors under full and partial observations. Given the partial observation O, the full observation is not uniquely determined because the remaining modalities are unobserved. The posterior under partial observation is the conditional mean of the posteriors under full observation that are compatible with the currently available modalities. This conditional mean does not characterize how much the posterior may vary across those full observations.

3.2 Uncertainty from Missing Modalities

Proposition 1 characterizes the posterior under partial observation as a conditional mean. We next consider the variability of the posterior under full observation given the current partial observation O. We define the total conditional variance of the posterior under full observation as uncertainty from missing modalities:

$$U_M(O) = \operatorname{tr} \operatorname{Cov}\left(\boldsymbol{\eta}_F(X) \mid O\right). \tag{3}$$

The scalar $U_M(O)$ measures how much the posterior under full observation may vary given the current partial observation. This definition focuses on uncertainty arising from missing modalities rather than class ambiguity in general. A diffuse posterior under partial observation does not necessarily imply a large $U_M(O)$ if the posterior changes little across compatible full observations.

Proposition 1 and the definition of conditional variance lead to the following result.

Proposition 2 (Uncertainty as Expected Squared Change). The uncertainty in Eq. 3 equals the expected squared change between the posteriors under full and partial observations conditioned on the current partial observation:

$$U_M(O) = \mathbb{E}\left[\|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2^2 \mid O\right]. \tag{4}$$

The same quantity can also be written as

$$U_M(O) = \mathbb{E}\left[\|\boldsymbol{\eta}_F(X)\|_2^2 \mid O\right] - \|\boldsymbol{\eta}_M(O)\|_2^2.$$

It satisfies $0 \le U_M(O) \le 1 - 1/C$.

The proof of Proposition 2 is provided in Appendix A.2. For the four-class brain tumor segmentation task considered in this work, $C = 4$, giving $0 \le U_M(O) \le 3/4$.

Eq. 4 provides a direct interpretation of uncertainty from missing modalities. If the posterior varies little across full observations compatible with the current partial observation, $U_M(O)$ is small. Larger posterior variation across these full observations corresponds to higher uncertainty from missing modalities. The uncertainty considered in this work is directly related to posterior changes associated with missing modalities rather than being determined solely by the concentration of the current posterior.

This distinction is consistent with the phenomenon illustrated in Fig. 1. Prediction confidence reflects the model's support for a particular class under the current observation. Prediction discrepancy reflects the change between predictions under full and partial observations. These quantities therefore describe different properties.

3.3 Observable Prediction Change

Eq. 4 expresses $U_M(O)$ as a conditional expectation over full observations compatible with the current partial observation O. This conditional expectation cannot be directly observed from an individual training sample. We first define the sample-level squared change between the class posteriors under full and partial observations:

$$Z = \|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2^2 . \tag{5}$$

The quantity Z represents the posterior change associated with a particular pair of full and partial observations and is not itself the uncertainty from missing modalities. Proposition 2 gives $\mathbb{E}[Z \mid O] = U_M(O)$. The uncertainty $U_M(O)$ is the conditional average of the sample-level posterior change given the current partial observation.

To show how Z can be used to learn $U_M(O)$, consider an arbitrary scalar function $u(O)$ that depends only on the current partial observation O. We define its population risk under squared loss as

$$\mathcal{R}(u) = \mathbb{E}\left[(u(O) - Z)^2\right]. \tag{6}$$

This gives the following result.

Proposition 3 (Optimality of Prediction Change Supervision under Squared Loss). Among all square-integrable O-measurable functions, the minimizer of the population risk in Eq. 6 is

$$u^*(O) = \mathbb{E}[Z \mid O] = U_M(O). \tag{7}$$

The proof of Proposition 3 is provided in Appendix A.3.

Proposition 3 shows that the posterior change Z for an individual sample is not required to equal its uncertainty from missing modalities. The sample-level quantity Z can serve as a supervision signal. When this quantity is learned over the data distribution using squared loss, the population risk is minimized by the uncertainty defined in Eq. 3. This establishes a direct statistical connection between the sample-level posterior change under full and partial observations and the theoretically defined uncertainty.

In practice, the theoretical class posteriors are not directly available during training. We use the predicted class probability vectors $\mathbf{q}_F(X)$ and $\mathbf{q}_M(O)$ as estimates of the corresponding theoretical posteriors. The practical quantity corresponding to Eq. 5 is

$$\widehat{Z} = \|\mathbf{q}_F(X) - \mathbf{q}_M(O)\|_2^2 . \tag{8}$$

When the predicted class probabilities approximate their corresponding class posteriors, $\widehat{Z}$ provides an observable approximation to the theoretical sample-level change Z. Appendix A.4 derives an upper bound on this approximation error in terms of the posterior estimation errors under full and partial observations. The predictions available during training therefore provide computable supervision for learning the uncertainty characterized in Eq. 7.

These results show that the prediction changes associated with missing modalities provide supervision for uncertainty learning. Sec. 4 develops the corresponding uncertainty representation and learning method.

4 Method

In multimodal brain tumor segmentation, different MRI modalities provide complementary tissue contrasts and pathological characteristics. With all modalities available, the model forms class probability predictions from the full observation. When one or more modalities are missing, predictions are formed from the partial observation and may differ from those under full observation. The theoretical analysis in Sec. 3 establishes a direct probabilistic connection between posterior variation under full and partial observations and uncertainty from missing modalities. This relationship forms the basis of the uncertainty representation and learning objective described below.

We develop an uncertainty learning framework for segmentation with missing modalities. We model the predicted class probabilities under partial observation with a Dirichlet distribution and estimate the concentration parameter from the decoded features. The resulting Dirichlet total variance provides a parameterized representation of uncertainty from missing modalities. We construct an uncertainty learning objective that matches the Dirichlet total variance to the prediction change between full and partial observations during training. The overall framework is illustrated in Fig. 2.

4.1 Dirichlet Uncertainty Modeling

To characterize uncertainty in the predicted class probabilities under partial observation, we use a Dirichlet distribution to model the prediction. Given the partial observation O, let $\mathbf{q}_M(O)$ denote the voxel-wise class probability vector predicted by the model. A lightweight concentration head estimates a positive concentration from the decoded feature $\mathbf{F}_M(O)$:

$$S_M(O) = \mathrm{Softplus}\left(h\left(\mathbf{F}_M(O)\right)\right), \qquad S_M(O) > 0. \tag{9}$$

The function $h(\cdot)$ denotes the concentration head implemented with a $1 \times 1 \times 1$ convolution. In our implementation, the concentration head receives a detached version of the decoded feature, so gradients from uncertainty learning do not propagate through the segmentation features.

The predicted class probability vector $\mathbf{q}_M(O)$ and concentration $S_M(O)$ parameterize the Dirichlet distribution:

$$\boldsymbol{\pi}_M(O) \mid O \sim \mathrm{Dir}\left(\boldsymbol{\alpha}_M(O)\right), \qquad \boldsymbol{\alpha}_M(O) = S_M(O)\mathbf{q}_M(O). \tag{10}$$

The predicted probabilities satisfy $\sum_{c=1}^{C} q_{M,c}(O) = 1$, so $S_M(O) = \sum_{c=1}^{C} \alpha_{M,c}(O)$ is the total concentration of the Dirichlet distribution. Its conditional mean is

$$\mathbb{E}\left[\boldsymbol{\pi}_M(O) \mid O\right] = \mathbf{q}_M(O).$$

The total variance of the Dirichlet distribution is

$$V_M(O) = \mathrm{tr}\,\mathrm{Cov}\left(\boldsymbol{\pi}_M(O) \mid O\right) = \frac{1 - \|\mathbf{q}_M(O)\|_2^2}{S_M(O) + 1}. \tag{11}$$

We use $V_M(O)$ as a parameterized representation of uncertainty from missing modalities. The uncertainty learning objective aligns this Dirichlet total variance with the prediction changes associated with missing modalities, connecting the model representation to the theoretical uncertainty defined in Sec. 3.2.

4.2 Uncertainty Learning from Prediction Changes

The theoretical analysis in Sec. 3 establishes the relationship between posterior changes under full and partial observations and uncertainty from missing modalities. During training, we use the observable prediction change $\widehat{Z}(X, O)$ defined in Eq. 8. The prediction change is computed independently at each spatial location and summed over the class dimension for voxel-wise segmentation.

The theoretical relationship in Sec. 3 is defined in terms of the true class posteriors, whereas the network uses predicted class probabilities as their estimates. We use the Brier score to encourage these predictions to approximate their corresponding posteriors. For a predicted class probability vector $\mathbf{q}(O)$, the class label Y and its one-hot representation $\mathbf{e}_Y$, the conditional Brier risk satisfies

$$\mathbb{E}_{p(Y|O)}\left[\|\mathbf{q}(O) - \mathbf{e}_Y\|_2^2\right] = \|\mathbf{q}(O) - \boldsymbol{\eta}_M(O)\|_2^2 + 1 - \|\boldsymbol{\eta}_M(O)\|_2^2. \tag{12}$$

The population Brier risk is minimized by the corresponding class posterior:

$$\operatorname*{argmin}_{\mathbf{q}} \mathbb{E}_{p(Y|O)}\left[\|\mathbf{q} - \mathbf{e}_Y\|_2^2\right] = \boldsymbol{\eta}_M(O). \tag{13}$$

The same property holds under full observation by replacing O and $\boldsymbol{\eta}_M(O)$ with X and $\boldsymbol{\eta}_F(X)$. The Brier loss encourages the predicted class probabilities under full and partial observations to approximate their corresponding class posteriors.

We train the Dirichlet total variance $V_M(O)$ defined in Sec. 4.1 using the observable prediction change:

$$\mathcal{L}_{\mathrm{PO}} = \mathbb{E}_{p(O)}\left[\mathbb{E}_{p(X|O)}\left[\left(V_M(O) - \widehat{Z}(X,O)\right)^2\right]\right]. \tag{14}$$

The observable prediction change is used as the regression target for learning the Dirichlet uncertainty representation, following the theoretical relationship established in Sec. 3.

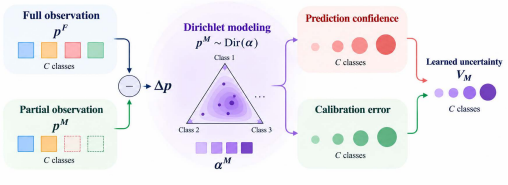


Figure 2: Overview of the proposed uncertainty learning framework. Under partial observation, the Dirichlet total variance V_M is learned from the prediction change between full and partial observations.

Final Objective. In addition to the uncertainty learning objective, we use the segmentation loss to supervise predictions under both full and partial observations. Prediction-level knowledge distillation transfers information from the prediction under full observation to that under partial observation. The Brier loss is applied to the predicted class probabilities under both observations. The final training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{seg}} + \beta\mathcal{L}_{\mathrm{Brier}} + \gamma\mathcal{L}_{\mathrm{KD}} + \delta\mathcal{L}_{\mathrm{PO}}. \tag{15}$$

$\mathcal{L}_{\mathrm{seg}}$ denotes the segmentation loss under full and partial observations, $\mathcal{L}_{\mathrm{KD}}$ denotes the prediction-level knowledge distillation loss from full to partial observation, $\mathcal{L}_{\mathrm{Brier}}$ denotes the Brier loss under both observations and $\mathcal{L}_{\mathrm{PO}}$ denotes the proposed uncertainty learning objective. The coefficients β, γ and δ control the corresponding loss terms.

5 Experiments

5.1 Segmentation Performance across Modality Combinations

Experimental Settings. To evaluate the effectiveness of our method for brain tumor segmentation with missing modalities, we conduct benchmark experiments on BraTS2020 (Bakas et al., 2017) and BraTS2024 (de Verdier et al., 2024) without any pretraining. Four modalities (T1, T1ce, T2, and FLAIR) are involved in both datasets. We use the Dice similarity coefficient and the 95th percentile Hausdorff distance (HD95) as evaluation metrics. For fair comparison, all methods use the same dataset split, with 80% of the data for training and 20% for testing. The comparison methods are reproduced using their publicly released implementations. All experiments are conducted on a single NVIDIA L40 GPU.

Segmentation Results. We compare our method with five widely used approaches for segmentation with missing modalities: RFNet (Ding et al., 2021), mmFormer (Zhang et al., 2022), M^{2}FTrans (Shi et al., 2024), IMS^{2}Trans (Zhang et al., 2024) and MSTKD (Zhu et al., 2025). MSTKD is a recent knowledge distillation method designed for segmentation with missing modalities. Tab. 1 presents the complete comparison results. In addition to the Dice evaluation reported in the main text, Appendix A.5 provides a comparison using HD95 to further evaluate boundary localization performance. Appendix A.9 further presents qualitative comparisons using 2D and 3D visualizations on BraTS2024.

Our method shows strong segmentation performance across both benchmarks. For WT, it achieves the best Dice score under every modality combination. The advantage becomes more pronounced when two or three modalities are missing. The performance gap becomes smaller when only one modality is missing. This trend is consistent with our motivation that uncertainty decreases as more modalities become available. For ET, our method falls outside the top two under a few modality combinations. These variations are limited to individual configurations. The method still achieves the highest average ET Dice score on BraTS2020. The ranking across individual modality combinations changes between the two benchmarks. Our method achieves either the best or second-best result under every modality combination on BraTS2024.

Table 1: Missing modality segmentation results on BraTS2020 and BraTS2024. For each modality combination, from left to right, the four circles represent T2, T1ce, T1 and FLAIR, respectively. ● and ○ denote available and missing modalities, respectively. The best results are shown in **bold**.

	Modality	●○○○	○●○○	○○●○	○○○●	●●○○	○●●○	○○●●	●○●○	●○○●	○●○●	○●●●	●○●●	●●○●	●●●○	●●●●	Avg.
BraTS2020																	
WT	RFNet	86.05	76.77	77.16	87.32	87.74	81.12	89.73	87.73	89.87	89.89	90.69	90.60	90.68	88.25	91.11	86.98
	mmFormer	83.88	75.80	75.11	86.84	86.57	79.71	88.71	86.10	89.22	88.82	89.40	89.50	89.70	86.90	89.85	85.74
	IMS^{2}Trans	84.35	72.91	70.83	86.50	86.51	75.58	88.94	86.64	89.43	88.16	88.96	89.99	89.93	87.27	90.27	85.08
	M^{2}FTrans	84.75	76.07	75.21	87.12	86.82	79.46	87.90	85.55	88.44	88.17	88.22	88.49	88.99	87.28	88.97	85.43
	MSTKD	84.39	69.67	60.38	81.70	85.59	72.96	85.42	84.34	89.49	89.49	89.77	88.71	90.57	85.27	90.70	83.23
	Ours	**87.24**	**79.87**	**78.75**	**90.05**	**88.17**	**81.90**	**91.12**	**88.22**	**91.58**	**91.18**	**91.41**	**91.62**	**91.64**	**88.27**	**91.74**	**88.18**
TC	RFNet	71.01	81.51	66.02	69.19	83.45	83.40	73.07	73.13	74.14	84.65	85.07	75.19	84.97	83.47	85.21	78.23
	mmFormer	69.50	82.83	63.54	68.02	84.27	84.21	72.33	72.11	72.87	83.82	84.90	74.22	84.42	84.91	84.85	77.79
	IMS^{2}Trans	66.39	80.14	56.09	62.15	84.94	81.06	67.80	67.20	69.58	83.69	83.30	69.88	84.57	84.67	84.28	75.05
	M^{2}FTrans	68.62	82.16	62.77	68.94	84.55	83.32	72.40	71.30	73.24	84.50	85.07	73.96	85.19	85.19	85.39	77.77
	MSTKD	69.29	73.92	51.64	61.60	81.17	75.81	70.57	69.47	71.83	82.29	83.24	72.44	82.20	81.47	82.17	73.94
	Ours	**72.15**	**84.99**	**68.87**	**71.46**	**86.02**	**85.68**	**75.95**	**73.61**	**74.78**	**86.47**	**86.78**	**75.69**	**86.14**	**86.59**	**86.41**	**80.11**
ET	RFNet	46.29	74.85	37.30	38.15	75.93	**78.01**	40.98	45.65	49.32	76.67	76.81	49.92	77.12	76.99	78.00	61.47
	mmFormer	42.56	**76.95**	31.80	39.61	76.09	76.60	41.48	43.94	44.98	76.27	76.72	46.81	75.21	76.62	75.67	60.09
	IMS^{2}Trans	44.25	74.90	30.77	37.54	77.50	76.73	41.95	44.12	**50.50**	78.85	79.46	**49.97**	77.85	78.40	**79.85**	61.51
	M^{2}FTrans	44.53	76.33	38.18	39.15	**78.06**	77.33	42.85	45.96	46.15	77.79	77.53	47.89	77.87	**79.66**	79.09	61.89
	MSTKD	45.21	73.44	31.73	38.24	74.25	71.66	43.15	44.91	46.97	74.50	74.97	47.32	73.41	74.12	74.52	59.23
	Ours	**46.76**	75.03	**39.29**	**41.73**	77.46	76.51	**45.83**	**46.21**	49.66	**79.00**	**80.63**	48.54	**79.50**	77.30	77.52	**62.73**
BraTS2024																	
WT	RFNet	73.38	67.57	68.12	81.22	76.83	70.83	83.09	76.27	83.33	83.87	84.22	83.83	84.50	77.23	84.56	78.59
	mmFormer	71.46	63.57	65.29	81.31	75.47	68.45	82.46	74.75	83.17	83.32	83.62	83.19	84.08	76.19	83.97	77.35
	IMS^{2}Trans	70.84	59.83	60.90	80.70	73.87	65.09	81.21	74.36	82.34	81.56	81.39	82.31	82.87	74.97	82.44	75.65
	M^{2}FTrans	68.19	58.98	60.07	76.71	73.41	64.55	79.87	72.49	80.31	81.06	81.77	81.05	82.22	74.19	82.35	74.48
	MSTKD	75.03	68.79	68.47	84.04	78.74	71.56	85.09	77.69	85.72	86.58	86.71	85.78	87.20	79.16	87.27	80.52
	Ours	**79.06**	**74.78**	**75.07**	**87.30**	**81.48**	**77.04**	**88.24**	**80.97**	**88.45**	**89.16**	**89.34**	**88.75**	**89.66**	**82.13**	**89.73**	**84.08**
TC	RFNet	29.66	57.12	**31.02**	24.87	62.21	61.54	33.42	34.47	32.53	60.36	63.52	35.32	63.52	63.76	64.01	47.82
	mmFormer	24.62	48.84	25.67	24.52	52.64	56.04	31.19	30.14	31.69	52.06	59.52	35.17	54.55	57.12	58.90	42.84
	IMS^{2}Trans	30.25	58.29	30.07	29.94	60.08	62.92	**40.50**	**41.24**	**41.25**	57.71	62.63	**43.16**	60.26	63.68	63.31	49.69
	M^{2}FTrans	28.93	58.81	27.07	24.58	62.99	62.86	30.81	31.89	31.70	62.81	64.96	33.84	63.11	63.61	64.41	47.49
	MSTKD	28.14	**64.74**	29.73	28.25	**65.25**	**64.78**	32.14	32.01	31.90	62.18	65.21	32.78	63.17	**66.54**	65.81	48.84
	Ours	**30.45**	60.92	30.67	**30.84**	63.50	63.43	35.35	34.69	35.20	**66.97**	**67.89**	35.79	**67.95**	64.54	**68.67**	**50.46**
ET	RFNet	26.30	58.05	27.92	21.93	63.19	62.95	29.92	31.20	29.47	61.73	64.45	32.15	64.45	66.14	65.80	47.04
	mmFormer	21.55	48.94	21.81	21.00	52.80	57.09	27.40	26.84	28.33	52.61	60.41	31.88	54.99	57.70	59.99	41.56
	IMS^{2}Trans	28.17	58.27	28.44	27.84	61.50	63.31	32.29	32.40	32.27	58.17	63.85	**40.08**	61.62	64.93	64.90	47.87
	M^{2}FTrans	25.11	60.43	21.85	21.20	64.81	64.19	26.63	28.15	27.62	64.67	67.43	29.92	65.39	67.23	67.24	46.79
	MSTKD	25.85	**65.32**	27.42	25.87	63.93	65.63	29.41	30.06	29.51	63.85	66.33	30.16	64.52	65.30	67.18	48.02
	Ours	**32.50**	61.09	**36.05**	**32.56**	**66.19**	**66.27**	**37.44**	**38.77**	**39.23**	**67.06**	**68.57**	33.13	**68.23**	**67.27**	**69.38**	**52.25**

5.2 Uncertainty Estimation under Missing Modalities

Uncertainty Visualization. To provide an intuitive view of how the model responds to uncertainty introduced by missing modalities, we visualize the estimated uncertainty under different concentra-

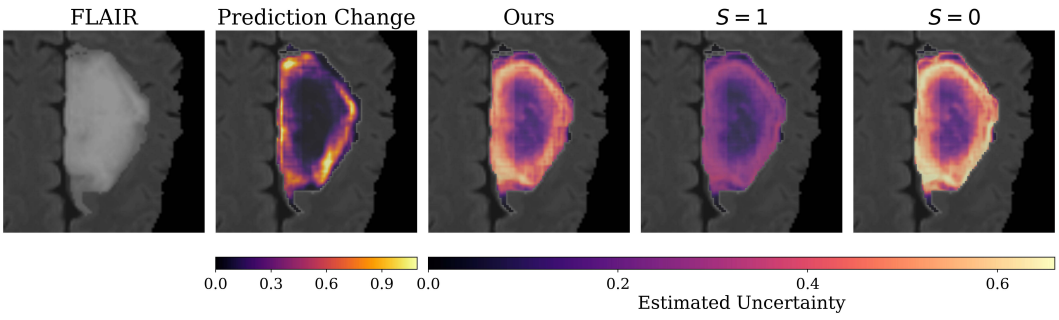


Figure 3: Uncertainty visualization with only FLAIR available. Brighter colors within the tumor region indicate higher prediction change or estimated uncertainty. Prediction change is computed using Eq. 8 and serves as an observable reference for the effect of missing modalities on the prediction. S denotes the concentration parameter in Eq. 11. We compare two settings, $S = 0$ and $S = 1$.

Table 2: Quantitative comparison of uncertainty estimation under different concentration settings.

Metric	$S = 0$	$S = 1$	$S = 2$	$S = 4$	$S = 8$	**Ours**
RMSE ↓	0.124897	0.121989	0.126221	0.131508	0.135883	**0.117539**
UCE ↓	0.015440	0.015535	0.018965	0.022484	0.025161	**0.012485**

tion settings. Fig. 3 shows the comparison. The uncertainty estimated by our method exhibits a spatial pattern broadly consistent with the prediction change. Regions with larger prediction changes tend to receive higher uncertainty responses. This indicates that the estimated uncertainty can reflect the effect of missing modalities to a reasonable extent. The results obtained with $S = 0$ and $S = 1$ show less spatial adaptation. According to Eq. 11, the estimated uncertainty depends on both the class probabilities and the concentration parameter. The learned $S_M(O)$ therefore allows the uncertainty response to adapt to the input.

Quantitative Evaluation. Tab. 2 compares the learned concentration with fixed concentration values $S = 0, 1, 2, 4, 8$ using RMSE and UCE. Our method achieves the lowest RMSE and the lowest UCE among all evaluated settings, showing that the learned uncertainty provides a more accurate and better calibrated characterization of prediction variation caused by missing modalities. These results support the benefit of learning an adaptive concentration parameter rather than using a fixed concentration. For completeness, the corresponding uncertainty calibration curves are provided in Appendix A.8.

5.3 Ablation Study

Tab. 3 presents the ablation study of $\mathcal{L}_{\mathrm{PO}}$ and $\mathcal{L}_{\mathrm{Brier}}$. The baseline retains the segmentation and knowledge distillation losses and removes L_{PO} and L_{Brier}. We further evaluate two settings that retain only one of the losses. Removing either loss reduces the mean segmentation performance of the complete model. The complete model achieves the strongest overall result among the evaluated settings.

Table 3: Ablation study of the proposed uncertainty learning objectives.

$\mathcal{L}_{\mathrm{PO}}$	$\mathcal{L}_{\mathrm{Brier}}$	WT	TC	ET	Mean
–	–	81.84	49.15	47.90	59.63
✓	–	83.82	48.94	49.67	60.81
–	✓	83.86	49.79	48.92	60.86
✓	✓	84.08	50.46	52.25	62.26

The two losses address different aspects of the learning problem. $\mathcal{L}_{\mathrm{Brier}}$ improves the learning of class probabilities. $\mathcal{L}_{\mathrm{PO}}$ encourages the estimated uncertainty to reflect the prediction change introduced in Sec. 3.3.

6 Conclusion

We studied uncertainty from missing modalities in multimodal MRI brain tumor segmentation. We first observed that prediction confidence can remain high even when predictions change under missing modalities. We then characterized this uncertainty as the total conditional variance of the class posterior under full observation given the current partial observation. We further established that this quantity equals the expected squared change between the class posteriors under full and partial observations. This analysis provides a theoretical basis for using prediction change as a supervision signal for uncertainty learning.

We developed a joint framework for segmentation and uncertainty estimation under partial observation. The framework combines prediction-level knowledge distillation, the Brier loss and a Dirichlet uncertainty model with a learned concentration parameter. Experiments on BraTS2020 and BraTS2024 show competitive segmentation performance across different combinations of missing modalities. Additional evaluations of boundary localization, uncertainty calibration, parameter sensitivity and qualitative comparison are provided in Appendix A. These evaluations complement the segmentation and uncertainty analyses in the main paper.

AI USE STATEMENT

Large language models (LLMs) were used solely to improve language clarity and readability. All scientific ideas, methodology, experiments, analyses and conclusions were developed and verified by the authors.

ETHICS STATEMENT

This study uses the BraTS2020 and BraTS2024 benchmark datasets and does not involve direct recruitment of human participants or access to personally identifiable information.

REPRODUCIBILITY STATEMENT

Implementation details and training settings are provided in the main paper and supplementary materials. Source code and evaluation scripts will be made available to facilitate reproduction of the reported results.

REFERENCES

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

Thomas Buddenkotte, Lorena Escudero Sanchez, Mireia Crispin-Ortuzar, Ramona Woitek, Cathal McCague, James D. Brenton, Ozan Oktem, Evis Sala, and Leonardo Rundo. Calibrating ensembles for scalable uncertainty quantification in deep learning-based medical image segmentation. *Computers in Biology and Medicine*, 163:107096, 2023.

Robin Camarasa, Daniel Bos, Jeroen Hendrikse, Paul Nederkoorn, Eline Kooi, Aad van der Lugt, and Marleen de Bruijne. Quantitative comparison of monte-carlo dropout uncertainty measures for multi-class segmentation. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pp. 32–41, 2020.

Yoonseok Choi, Mohammed A. Al-Masni, Kyu-Jin Jung, Roh-Eul Yoo, Seong-Yeong Lee, and Dong-Hyun Kim. A single stage knowledge distillation network for brain tumor segmentation on limited MR image modalities. *Computer Methods and Programs in Biomedicine*, 240:107644, 2023.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, et al. The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI. *arXiv preprint arXiv:2405.18368*, 2024.

Yueqin Diao, Huihui Fang, Hanyi Yu, Fan Li, and Yanwu Xu. Multimodal invariant feature prompt network for brain tumor segmentation with missing modalities. *Neurocomputing*, 616:128847, 2025.

Yuhang Ding, Xin Yu, and Yi Yang. 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.

Mohammad Hamghalam, Alejandro F. Frangi, Baiying Lei, and Amber L. Simpson. Modality completion via gaussian process prior variational autoencoders for multi-modal glioma segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pp. 442–452, 2021.

Ling Huang, Su Ruan, Yucheng Xing, and Mengling Feng. A review of uncertainty quantification in medical image analysis: Probabilistic and non-probabilistic methods. *Medical Image Analysis*, 97:103223, 2024.

Seung-wan Jeong, Hwan-ho Cho, Junmo Kwon, and Hyunjin Park. Region-of-interest attentive heteromodal variational encoder-decoder for segmentation with missing modalities. In *Proceedings of the Asian Conference on Computer Vision*, pp. 132–148, 2022.

Haitao Li, Ziyu Li, Yiheng Mao, Zhengyao Ding, and Zhengxing Huang. DC-Seg: Disentangled contrastive learning for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 138–148, 2025a.

Hao Li, Yang Nan, Javier Del Ser, and Guang Yang. Region-based evidential deep learning to quantify uncertainty and improve robustness of brain tumor segmentation. *Neural Computing and Applications*, 35(30):22071–22085, 2023.

Yuzhu Li, An Sui, Fuping Wu, and Xiahai Zhuang. Uncertainty-supervised interpretable and robust evidential segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 649–658, 2025b.

Zhiyuan Li, Yafei Zhang, Huafeng Li, Yi Chai, and Yushi Yang. Deformation-aware and reconstruction-driven multimodal representation learning for brain tumor segmentation with missing modalities. *Biomedical Signal Processing and Control*, 91:106012, 2024.

Tianyi Liu, Haochuan Jiang, and Kaizhu Huang. KMD: Koopman multi-modality decomposition for generalized brain tumor segmentation under incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15663–15671, 2025a.

Weide Liu, Jingwen Hou, Xiaoyang Zhong, Huijing Zhan, Jun Cheng, Yuming Fang, and Guanghui Yue. Improving multi-modal brain tumor segmentation via pre-training and knowledge distillation based post-training. *Neurocomputing*, 640:130318, 2025b.

Alireza Mehrtash, William M. Wells, Clare M. Tempany, Purang Abolmaesumi, and Tina Kapur. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Transactions on Medical Imaging*, 39(12):3868–3878, 2020.

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024, 2015.

Chengjian Qiu, Yuqing Song, Yi Liu, Yan Zhu, Kai Han, Victor S. Sheng, and Zhe Liu. MMMViT: Multiscale multimodal vision transformer for brain tumor segmentation with missing modalities. *Biomedical Signal Processing and Control*, 90:105827, 2024.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pp. 3179–3189, 2018.

Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. M^2FTrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):379–390, 2024.

Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. Classifier calibration: A survey on how to assess and improve predicted class probabilities. *Machine Learning*, 112(9):3211–3260, 2023.

Hsienchih Ting and Manhua Liu. Multimodal transformer of incomplete MRI data for brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):89–99, 2024.

Zirui Wang and Yi Hong. A2FSeg: Adaptive multi-modal fusion network for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pp. 673–681, 2023.

Heran Yang, Jian Sun, and Zongben Xu. Learning unified hyper-network for multi-modal MR image synthesis and tumor segmentation with missing modalities. *IEEE Transactions on Medical Imaging*, 42(12):3678–3689, 2023.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: Arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 424–433, 2024.

Dongsong Zhang, Changjian Wang, Tianhua Chen, Weidao Chen, and Yiqing Shen. Scalable swin transformer network for brain tumor segmentation from incomplete MRI modalities. *Artificial Intelligence in Medicine*, 149:102788, 2024.

Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmFormer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 107–117, 2022.

Yidong Zhao, Changchun Yang, Artur Schweidtmann, and Qian Tao. Efficient bayesian uncertainty estimation for nnU-Net. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 535–544, 2022.

Shenhai Zheng, Xi Gao, Xue Tang, and Ziyun Wang. Collaborative encoding attributed hybrid attention knowledge distillation for missing-modality brain tumor segmentation. *Biomedical Signal Processing and Control*, 113:109220, 2026.

Tongxue Zhou, Su Ruan, and Haigen Hu. A literature survey of MR-based brain tumor segmentation with missing modalities. *Computerized Medical Imaging and Graphics*, 104:102167, 2023.

Shenghao Zhu, Yifei Chen, Weihong Chen, Yuanhan Wang, Chang Liu, Shuo Jiang, Feiwei Qin, and Changmiao Wang. Bridging the gap in missing modalities: Leveraging knowledge distillation and style matching for brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 95–106, 2025.

Shenghao Zhu, Yifei Chen, Weihong Chen, Shuo Jiang, Guanyu Zhou, Yuanhan Wang, Feiwei Qin, Changmiao Wang, and Qiyuan Tian. No modality left behind: Adapting to missing modalities via knowledge distillation for brain tumor segmentation. *Medical Image Analysis*, 112:104108, 2026.

Ke Zou, Xuedong Yuan, Xiaojing Shen, Meng Wang, and Huazhu Fu. TBraTS: Trusted brain tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 503–513, 2022.

Ke Zou, Yidi Chen, Ling Huang, Nan Zhou, Xuedong Yuan, Xiaojing Shen, Meng Wang, Rick Siow Mong Goh, Yong Liu, Yih Chung Tham, and Huazhu Fu. Toward reliable medical image segmentation by modeling evidential calibrated uncertainty. *IEEE Transactions on Cybernetics*, 55(12):5975–5988, 2025.

A Appendix

This appendix provides detailed proofs of Propositions 1–3 in Sec. 3 and further analyzes the approximation error between the observable prediction change $\widehat{Z}$ constructed from network predictions and the theoretical sample-level change Z. We follow the notation introduced in Sec. 3. Additional experimental results are also provided to complement and support the analyses presented in the main paper.

A.1 Proof of Proposition 1

Proof. Since the class posterior is a C-dimensional probability vector, it is sufficient to establish the result for each class $c \in \{1, \dots, C\}$. Define the class indicator $I_c = \mathbf{1}\{Y = c\}$. By the relation between conditional probability and conditional expectation, $\eta_{F,c}(X) = \mathbb{E}[I_c \mid X]$.

Under the conditional independence assumption $Y \perp M \mid X$, the modality availability pattern provides no additional class information once the full observation X is given. The assumption gives

$$P(Y = c \mid X, M) = P(Y = c \mid X).$$

By the definition of I_c,

$$E[I_c \mid X, M] = P(Y = c \mid X, M) = \eta_{F,c}(X).$$

Since $O = (X_M, M)$ is a deterministic function of (X, M), $\sigma(O) \subseteq \sigma(X, M)$. For any integrable random variable W, the tower property gives

$$\mathbb{E}[\mathbb{E}[W \mid X, M] \mid O] = \mathbb{E}[W \mid O].$$

Taking $W = I_c$, we obtain

$$\begin{aligned}\mathbb{E}[\eta_{F,c}(X) \mid O] &= \mathbb{E}[\mathbb{E}[I_c \mid X, M] \mid O] \\ &= \mathbb{E}[I_c \mid O] \\ &= P(Y = c \mid O) \\ &= \eta_{M,c}(O).\end{aligned}$$

Since this relation holds for every class c, combining all components gives

$$\boldsymbol{\eta}_M(O) = \mathbb{E}[\boldsymbol{\eta}_F(X) \mid O]\,.$$

Thus, given the current partial observation, the posterior under partial observation is the conditional average of the posteriors associated with full observations that are compatible with the currently available modalities. This proves Proposition 1.

A.2 Proof of Proposition 2

Proof. Recall that uncertainty from missing modalities is defined in Eq. 3 as $U_M(O) = \operatorname{tr}\operatorname{Cov}(\boldsymbol{\eta}_F(X) \mid O)$. By definition,

$$\operatorname{Cov}(\boldsymbol{\eta}_F(X) \mid O) = \mathbb{E}\Big[(\boldsymbol{\eta}_F(X) - \mathbb{E}[\boldsymbol{\eta}_F(X) \mid O])\,(\boldsymbol{\eta}_F(X) - \mathbb{E}[\boldsymbol{\eta}_F(X) \mid O])^\top \;\Big|\; O\Big]\,.$$

Substituting $\mathbb{E}[\eta_F(X) \mid O] = \eta_M(O)$ from Proposition 1 into the covariance gives

$$\operatorname{Cov}(\boldsymbol{\eta}_F(X) \mid O) = \mathbb{E}\Big[(\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O))\,(\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O))^\top \;\Big|\; O\Big]\,.$$

For any vector $\mathbf{d}$, $\operatorname{tr}(\mathbf{d}\mathbf{d}^\top) = \|\mathbf{d}\|_2^2$. Using this identity together with the linearity of the trace and conditional expectation yields

$$\begin{aligned}U_M(O) &= \operatorname{tr}\operatorname{Cov}(\boldsymbol{\eta}_F(X) \mid O) \\ &= \mathbb{E}\Big[\|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2^2 \mid O\Big]\,.\end{aligned}$$

This proves the conditional squared-change characterization in Eq. 4.

Expanding the squared norm gives

$$U_M(O) = \mathbb{E}\big[\|\boldsymbol{\eta}_F(X)\|_2^2 \mid O\big] + \|\boldsymbol{\eta}_M(O)\|_2^2 - 2\boldsymbol{\eta}_M(O)^\top \mathbb{E}[\boldsymbol{\eta}_F(X) \mid O]\,.$$

Applying Proposition 1 again,

$$U_M(O) = \mathbb{E}\big[\|\boldsymbol{\eta}_F(X)\|_2^2 \mid O\big] - \|\boldsymbol{\eta}_M(O)\|_2^2.$$

We next establish the bounds. Positive semidefiniteness of the covariance matrix gives $U_M(O) \geq 0$. The simplex constraints on $\eta_F(X)$ give

$$\|\boldsymbol{\eta}_F(X)\|_2^2 = \sum_{c=1}^{C} \eta_{F,c}^2(X) \leq \left(\sum_{c=1}^{C} \eta_{F,c}(X)\right)^2 = 1,$$

which implies $\mathbb{E}[\|\boldsymbol{\eta}_F(X)\|_2^2 \mid O] \leq 1$.

For the posterior under partial observation, the Cauchy–Schwarz inequality gives

$$1 = \left(\sum_{c=1}^{C} \eta_{M,c}(O) \right)^2 \leq C \sum_{c=1}^{C} \eta_{M,c}^2(O).$$

Hence, $\|\boldsymbol{\eta}_M(O)\|_2^2 \geq 1/C$. Combining these bounds gives

$$0 \leq U_M(O) \leq 1 - \frac{1}{C}.$$

For the four-class brain tumor segmentation setting considered in this work

$$0 \leq U_M(O) \leq \frac{3}{4}.$$

This completes the proof of Proposition 2.

A.3 Proof of Proposition 3

Proof. Recall the sample-level posterior change defined in Eq. 5,

$$Z = \|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2^2 .$$

Since both posterior vectors lie on the probability simplex, Z is bounded and therefore square-integrable.

For any square-integrable O-measurable scalar function $u(O)$, the population risk in Eq. 6 can be written using the law of total expectation as

$$\mathcal{R}(u) = \mathbb{E}\big[\mathbb{E}\big[(u(O) - Z)^2 \mid O\big]\big] .$$

For a fixed value of O, let a denote a candidate scalar prediction and define the corresponding conditional risk as

$$r_O(a) = \mathbb{E}\big[(a - Z)^2 \mid O\big] .$$

Expanding the square gives

$$\begin{aligned} r_O(a) &= \mathbb{E}\big[a^2 - 2aZ + Z^2 \mid O\big] \\ &= a^2 - 2a\,\mathbb{E}[Z \mid O] + \mathbb{E}[Z^2 \mid O]. \end{aligned}$$

For a given O, both $\mathbb{E}[Z \mid O]$ and $\mathbb{E}[Z^2 \mid O]$ are constant with respect to a. Differentiating the conditional risk with respect to a gives

$$\frac{\partial r_O(a)}{\partial a} = 2a - 2\mathbb{E}[Z \mid O].$$

Setting the derivative to zero yields $a^* = \mathbb{E}[Z \mid O]$. Moreover, $\partial^2 r_O(a)/\partial a^2 = 2 > 0$, so the conditional risk is strictly convex in a and this solution is its unique global minimizer.

Since the minimization holds conditionally for almost every O, the O-measurable function that minimizes the population risk is

$$u^*(O) = \mathbb{E}[Z \mid O].$$

By Proposition 2,

$$\begin{aligned} \mathbb{E}[Z \mid O] &= \mathbb{E}\Big[\|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2^2 \mid O\Big] \\ &= U_M(O). \end{aligned}$$

Therefore,

$$u^*(O) = \mathbb{E}[Z \mid O] = U_M(O).$$

This result clarifies the role of the sample-level quantity Z. An individual value of Z need not equal $U_M(O)$. When the sample-level changes are regressed using squared loss, the population risk is minimized by their conditional expectation given the current partial observation, which is exactly the uncertainty from missing modalities defined in Sec. 3.2. This completes the proof of Proposition 3.

A.4 Approximation Error Bound for the Observable Prediction Change

The theoretical quantity Z depends on the true class posteriors $\eta_F(X)$ and $\eta_M(O)$. These posteriors are not directly available during training. We use the predicted class probability vectors $q_F(X)$ and $q_M(O)$ to construct the observable prediction change in Eq. 8,

$$\widehat{Z} = \|\mathbf{q}_F(X) - \mathbf{q}_M(O)\|_2^2 .$$

We show below that the discrepancy between $\widehat{Z}$ and the theoretical sample-level change Z is bounded by the errors in estimating the class posteriors under full and partial observations.

To simplify notation, let $a = \|\mathbf{q}_F(X) - \mathbf{q}_M(O)\|_2$ and $b = \|\boldsymbol{\eta}_F(X) - \boldsymbol{\eta}_M(O)\|_2$. Then $\widehat{Z} = a^2$ and $Z = b^2$, and

$$|\widehat{Z} - Z| = |a^2 - b^2| = |a - b|(a + b).$$

By the reverse triangle inequality,

$$\begin{aligned} |a - b| &\le \|(\mathbf{q}_F - \mathbf{q}_M) - (\boldsymbol{\eta}_F - \boldsymbol{\eta}_M)\|_2 \\ &= \|(\mathbf{q}_F - \boldsymbol{\eta}_F) - (\mathbf{q}_M - \boldsymbol{\eta}_M)\|_2 \\ &\le \|\mathbf{q}_F - \boldsymbol{\eta}_F\|_2 + \|\mathbf{q}_M - \boldsymbol{\eta}_M\|_2, \end{aligned}$$

where the arguments have been omitted for readability.

For any two C-dimensional probability vectors $\mathbf{p}$ and $\mathbf{r}$, $\|\mathbf{p}\|_2^2 \le 1$, $\|\mathbf{r}\|_2^2 \le 1$ and $\mathbf{p}^\top \mathbf{r} \ge 0$.

$$\|\mathbf{p} - \mathbf{r}\|_2^2 = \|\mathbf{p}\|_2^2 + \|\mathbf{r}\|_2^2 - 2\mathbf{p}^\top \mathbf{r} \le 2.$$

The Euclidean distance between any two probability vectors on the simplex is at most $\sqrt{2}$. Since $\mathbf{q}_F$, $\mathbf{q}_M$, $\boldsymbol{\eta}_F$ and $\boldsymbol{\eta}_M$ are all probability vectors, $a \le \sqrt{2}$ and $b \le \sqrt{2}$, so $a + b \le 2\sqrt{2}$.

Combining the above inequalities gives

$$|\widehat{Z} - Z| \le 2\sqrt{2}\,(\|\mathbf{q}_F - \boldsymbol{\eta}_F\|_2 + \|\mathbf{q}_M - \boldsymbol{\eta}_M\|_2)\,.$$

Therefore, the approximation error between the observable prediction change and the theoretical sample-level change is bounded by the posterior estimation errors under full and partial observations.

Proposition 2 gives $U_M(O) = E[Z \mid O]$. The difference between $E[\widehat{Z} \mid O]$ and $U_M(O)$ can then be written as

$$\begin{aligned} \Big|\mathbb{E}[\widehat{Z} \mid O] - U_M(O)\Big| &= \Big|\mathbb{E}[\widehat{Z} - Z \mid O]\Big| \\ &\le \mathbb{E}\Big[|\widehat{Z} - Z| \mid O\Big]\,. \end{aligned}$$

Applying the previous bound yields

$$\Big|\mathbb{E}[\widehat{Z} \mid O] - U_M(O)\Big| \le 2\sqrt{2}\,\mathbb{E}[\|\mathbf{q}_F - \boldsymbol{\eta}_F\|_2 + \|\mathbf{q}_M - \boldsymbol{\eta}_M\|_2 \mid O]\,.$$

The bound links the approximation error to the posterior estimation errors. As these errors decrease, $\widehat{Z}$ approaches Z. Its conditional expectation approaches $U_M(O)$.

An individual value of $\widehat{Z}$ is not required to equal $U_M(O)$. The observable prediction change instead approximates the theoretical sample-level quantity Z, whose conditional expectation is the uncertainty from missing modalities. This provides the computable supervision used by the uncertainty learning objective in Sec. 4.2.

A.5 Boundary Localization Performance

We evaluate boundary localization performance using HD95. As shown in Tab. 4, our method achieves the lowest mean HD95 on both BraTS2020 and BraTS2024. It also obtains the lowest WT and TC HD95 on BraTS2020 while MSTKD gives a lower ET HD95. On BraTS2024, our method achieves the lowest HD95 for all three regions. These results demonstrate more accurate boundary localization across different missing-modality settings.

Table 4: Comparison of average HD95 results on BraTS2020 and BraTS2024 across all 15 modality combinations. Lower values indicate better performance.

BraTS2020	WT	TC	ET	Mean	BraTS2024	WT	TC	ET	Mean
RFNet	8.42	8.72	6.66	7.93	RFNet	11.91	10.63	10.11	10.89
mmFormer	7.71	8.09	6.11	7.30	mmFormer	9.55	9.95	9.38	9.62
MSTKD	6.46	7.08	5.54	6.36	MSTKD	6.67	8.50	8.15	7.77
Ours	5.73	6.51	5.74	5.99	Ours	4.97	7.05	6.75	6.26

A.6 Sensitivity Analysis

The coefficients β and δ control the contributions of $\mathcal{L}_{\text{Brier}}$ and $\mathcal{L}_{\text{PO}}$, respectively. As shown in Tab. 5, the performance varies to some extent under different combinations of the two coefficients and the effect is not identical across regions and datasets. On BraTS2020, the balanced setting $\beta = \delta = 1$ provides the strongest overall segmentation performance. On BraTS2024, several alternative settings improve individual regions, but no single alternative consistently performs better across all regions. We therefore use $\beta = \delta = 1$ in the main experiments since it provides the most consistent performance across the two datasets.

Table 5: Sensitivity analysis of β and δ on BraTS2020 and BraTS2024. The two values in each setting denote β and δ, respectively.

BraTS2020	$(1,1)$	$(0.5,0.5)$	$(0.5,2)$	$(2,0.5)$	BraTS2024	$(1,1)$	$(0.5,0.5)$	$(0.5,2)$	$(2,0.5)$
WT	88.18	87.05	87.06	86.96	WT	84.08	83.49	83.57	83.58
TC	80.11	75.86	75.35	75.41	TC	50.46	51.07	51.70	50.88
ET	62.73	62.25	61.89	61.96	ET	52.25	50.00	50.77	49.80

A.7 Computational Efficiency Analysis

As shown in Tab. 6, our method maintains a relatively small model size and low computational cost compared with the competing methods. It requires only 4.70M parameters and 149.05 GFLOPs, and achieves efficient inference with a throughput of 3.02 cases/s. Although the training time is not the shortest, the proposed method achieves the best average Dice scores for WT, TC and ET on BraTS2024. These results indicate that the proposed uncertainty modeling introduces limited computational overhead while preserving strong segmentation performance.

Table 6: Computational efficiency and segmentation performance on BraTS2024.

Method	Parameters (M)	GFLOPs	Training Time/Epoch (s)	Inference Time/Case (s)	Throughput (case/s)	WT	TC	ET
RFNet	8.99	420.89	218.21	0.8026	1.2459	78.59	47.82	47.10
mmFormer	36.65	130.19	810.92	0.1480	6.7585	77.35	42.84	41.56
M^{2}FTrans	45.43	433.95	243.23	1.4542	0.6876	74.48	47.49	46.79
IMS^{2}Trans	3.78	167.07	309.51	1.0508	0.9517	75.65	49.69	47.87
Ours	4.70	149.05	338.83	0.3307	3.0236	84.08	50.46	52.25

A.8 Calibration Analysis under Missing Modalities

We further examine calibration under missing modalities from two perspectives. Fig. 4 compares the estimated uncertainty with the corresponding prediction variation. Fixed concentration values lead to systematic underestimation or overestimation in different uncertainty ranges. At low uncertainty levels, $S = 0$ remains relatively close to the perfect calibration line. This suggests that only limited adjustment is needed in this regime. The deviation of fixed concentration values becomes more evident as uncertainty increases. Our learned adaptive concentration follows the perfect calibration line more closely in this region. This result shows that the proposed uncertainty modeling is particularly beneficial when the effect of missing modalities becomes more pronounced.

We also evaluate the predictive calibration of $\mathbf{q}_M$ using ECE, SCE and TACE. Tab. 7 shows low calibration errors across the 14 incomplete modality settings. The macro average ECE and SCE are

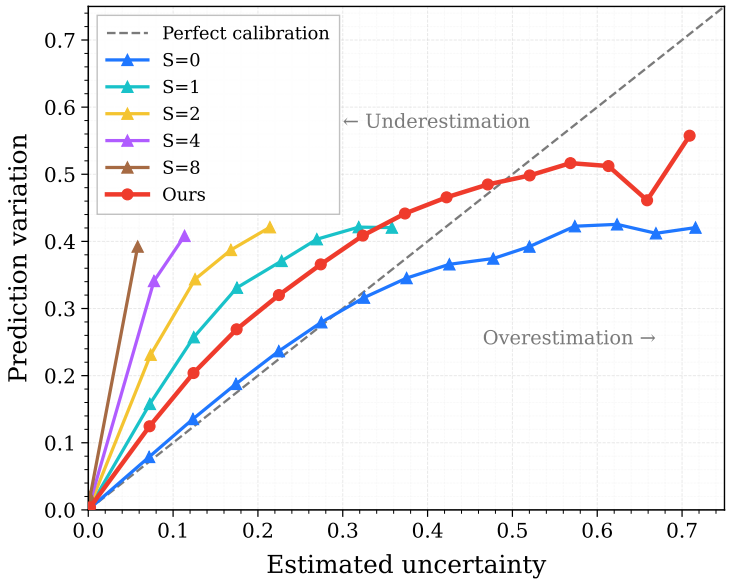


Figure 4: Uncertainty calibration under missing modalities. The dashed diagonal denotes perfect calibration. Curves above the diagonal indicate uncertainty underestimation. Curves below the diagonal indicate uncertainty overestimation. Ours uses the learned adaptive concentration. The fixed baselines use $S \in \{0, 1, 2, 4, 8\}$.

Table 7: Predictive calibration under different missing-modality settings on BraTS2020. The four symbols in each column correspond to FLAIR, T1ce, T1 and T2, respectively. A filled circle denotes an available modality and an open circle denotes a missing modality.

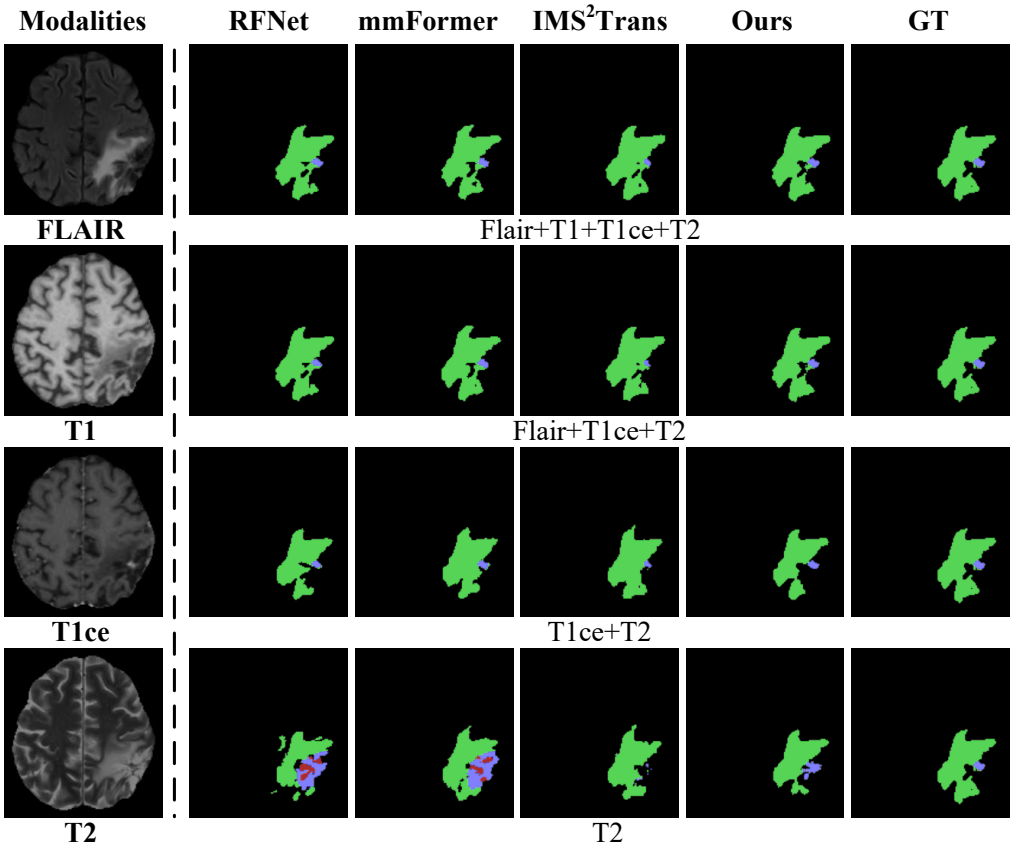
Metric	●○○○	○●○○	○○●○	○○○●	●●○○	●○●○	●○○●	○●●○	○●○●	○○●●	●●●○	●●○●	●○●●	○●●●	Avg.
ECE	0.0113	0.0098	0.0127	0.0103	0.0063	0.0102	0.0090	0.0092	0.0066	0.0097	0.0063	0.0066	0.0095	0.0069	0.0089
SCE	0.0105	0.0098	0.0129	0.0104	0.0065	0.0097	0.0091	0.0092	0.0069	0.0097	0.0063	0.0064	0.0092	0.0068	0.0088
TACE	0.0991	0.0642	0.0810	0.0851	0.0648	0.0934	0.0862	0.0639	0.0603	0.0828	0.0646	0.0645	0.0887	0.0614	0.0757

0.0089 and 0.0088. The corresponding TACE is 0.0757. These results indicate that the predicted class probabilities remain well calibrated under partial observation.

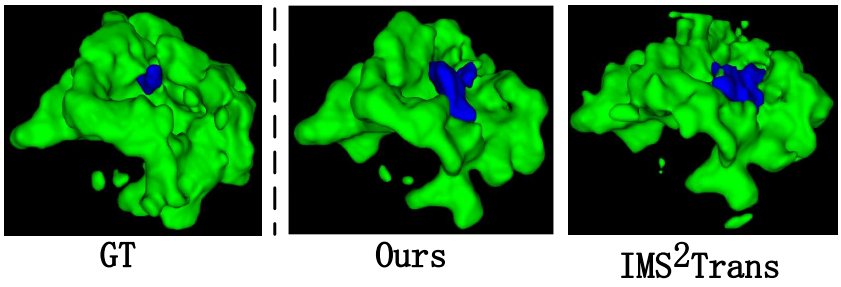
A.9 Qualitative Comparison on BraTS2024

As shown in Fig. 5(a), the differences among methods are relatively small when more modalities are available. The differences become more evident as fewer modalities remain available. With only T2 available, RFNet and mmFormer produce false positive regions. IMS^{2}Trans avoids most of these regions but misses part of the smaller tumor region. Our method maintains a stable segmentation under severe modality absence. Some degradation is still observed. This degradation is consistent with the loss caused by missing modalities and the difficulty of post-treatment cases in BraTS2024.

Fig. 5(b) provides a 3D comparison between our method and IMS^{2}Trans under the T2 setting. The prediction from our method follows the overall tumor morphology of the ground truth. The local shape is also retained in several regions. The prediction from IMS^{2}Trans shows some differences from the ground truth in these regions.



(a) Comparison under four representative modality settings.



(b) 3D comparison under the T2 setting.

Figure 5: Visual comparison on BraTS2024.