# Set-Inclusive Uncertainty Modeling for Robust Brain Tumor Segmentation

Seunghun Baek*, Jihwan Park*, Jaeyoon Sim, Hoseok Lee, Seungjoo Lee, and Won Hwa Kim

Pohang University of Science and Technology, Pohang, South Korea
{habaek4, pjh58110, simjy98, hslee0608, seungjoo612, wonhwa}@postech.ac.kr

**Abstract.** Multimodal MRI is essential for accurate brain tumor segmentation. However, acquiring all modalities at inference is often challenging in practice, which causes intrinsic uncertainty due to unavoidable information loss. Without modeling this uncertainty, existing methods encode incomplete evidence into deterministic representations that appear plausible but lack reliability. In this regime, we propose a probabilistic representation framework that models representations as Gaussian distributions, where their mean captures task information and their variance measures uncertainty from missing evidence. To make variance reflect information deficiency, we regularize the mean from each partial configuration toward its full-modality counterpart, while scaling the variance with the discrepancy between their aligned means. We further introduce a set-inclusive strategy that exploits the hierarchical structure of modality subsets and enforces an ordering constraint to maintain their consistent uncertainty relationships. Extensive experiments on BraTS 2018 and 2020 demonstrate that our approach offers superior performance over baselines across diverse missing-modality scenarios. Code and model checkpoint are available at https://github.com/atlas-sky/SIUM.

**Keywords:** Brain Tumor Segmentation · Probabilistic Representation

## 1 Introduction

Brain tumor segmentation is critical for assessing disease progression and treatment planning [3,5,17]. In clinical practice, four standard magnetic resonance imaging (MRI) modalities are utilized: T1, T1c, T2, and FLAIR [10,11], which provide complementary information for delineating tumor subregions, including whole tumor (WT), tumor core (TC), and enhancing tumor (ET) [4,6,11,17]. Since no single modality fully captures every tumor subregion, accurate delineation requires the joint integration of all modalities [12,18]. However, acquiring all of them is often impractical due to cost and patient burden [1,2]. As brain tumor segmentation relies on the complementary evidence from multiple modalities, when a modality goes missing, its unique information becomes inaccessible.

* S. Baek and J. Park contributed equally to this paper.

To address potential incompleteness, previous works [8,14,15,21] train their frameworks to be robust across missing-modality scenarios. They encode each modality configuration into a deterministic embedding to represent task-relevant information. While the absence introduces unavoidable information loss in the resulting embedding, deterministic approaches still update their prediction models without considering informational incompleteness (i.e., uncertainty). Consequently, the models may rely on hallucinated features that lack evidence.

In this regime, we model each configuration as a Gaussian distribution, where the mean encodes task-relevant information and the variance captures uncertainty. This design acknowledges that missing modalities cause intrinsic information deficiency, which should be reflected as uncertainty rather than concealed within a deterministic embedding. To make the variance encode uncertainty, we exploit the hierarchical structure of modality subsets, where smaller sets are nested within larger ones. Within this set-inclusive hierarchy, the variance is encouraged to increase when the mean of a subset deviates from its full-modality counterpart, since deviation indicates missing evidence. It is further regularized to remain higher for subsets than for their supersets, which reflects their reduced informational completeness. Through uncertainty guidance, the model enables reliable predictions that prioritize well-supported evidence across configurations.

**Our main contributions** are summarized as follows: **1)** We propose a probabilistic framework that integrates set-inclusive hierarchy with uncertainty guidance to model intrinsic uncertainty and emphasize reliable cues. **2)** We provide a theoretical analysis of how uncertainty influences task-representation optimization and empirical validation of the learned uncertainty to support the need for probabilistic embeddings. **3)** Extensive experiments on BraTS 2018 and 2020 demonstrate superior performance under diverse missing-modality scenarios.

**Related works on Brain Tumor Segmentation.** Early approaches adopt imputation methods that synthesize absent modalities [7,9,19]. However, these methods require additional generative models, which are computationally heavy and time-consuming. Recent works therefore have shifted toward imputation-free strategies. In particular, RFNet [8] introduces a region-aware fusion module that assigns adaptive modality importance across tumor regions. mmFormer [21] improves it using inter- and intra-modal Transformer [20] to model local and global context. M$^3$AE [15] mitigates missing-modality effects by modality- and patch-level masking and self-distillation. DC-Seg [14] disentangles modality-invariant and -specific representations via contrastive learning for anatomical consistency.

## 2 Methods

**General problem setting.** Given a sample $x_i$ (i.e., subject) with $M$ modalities, we denote its $m$-th modality input (e.g., T1 scan) as $x_i^{(m)}$ for $m = 1, \ldots, M$. Each modality input $x_i^{(m)}$ is encoded by a modality-specific encoder $\mathcal{E}^{(m)}$ to produce an embedding $e_i^{(m)} = \mathcal{E}^{(m)}(x_i^{(m)})$. To simulate missing-modality scenarios during training, a Bernoulli indicator $\delta_i^{(m)} \in \{0, 1\}$ is adopted to mask the

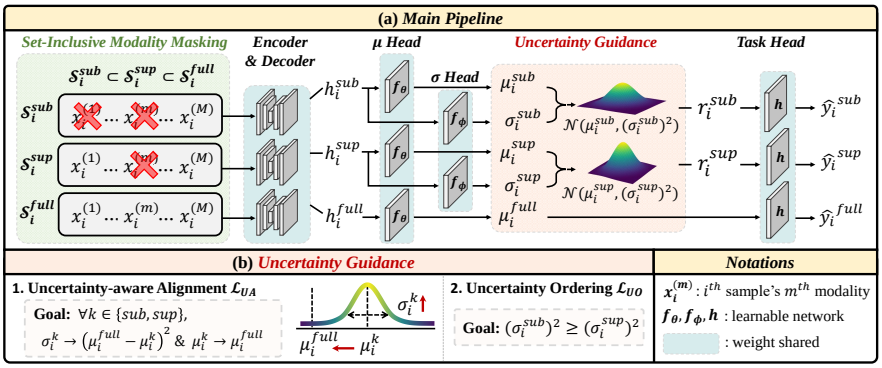


**Fig. 1.** Overall framework. **(a)** For each sample $x_i$, three modality sets $\mathcal{S}_i^{\text{sub}} \subset \mathcal{S}_i^{\text{sup}} \subset \mathcal{S}_i^{\text{full}}$ are constructed and then processed independently. Subset configurations are modeled probabilistically as $r_i^k \sim \mathcal{N}(\mu_i^k, (\sigma_i^k)^2)$, whereas the full configuration serves as a deterministic anchor as $\mu_i^{\text{full}}$. **(b)** $\mathcal{L}_{\text{UA}}$ aligns each subset mean $\mu_i^k$ with the full-modality anchor while enabling $\sigma_i^k$ to reflect their discrepancy. $\mathcal{L}_{\text{UO}}$ further enforces an ordering constraint such that smaller subsets maintain higher uncertainty than their supersets.

modality embedding, resulting in $\delta_i^{(m)} e_i^{(m)}$. A shared decoder $\mathcal{D}$ then aggregates the set of masked embeddings $\{\delta_i^{(m)} e_i^{(m)}\}_{m=1}^M$ to obtain a task-specific representation $z_i = \mathcal{D}(\{\delta_i^{(m)} e_i^{(m)}\}_{m=1}^M)$. While incomplete modalities inherently imply a high degree of uncertainty, existing methods [8,14,15,21] typically maintain $z_i$ as a deterministic embedding. By directly passing $z_i$ to a task head $h$ to obtain the prediction $\hat{y}_i = h(z_i)$, such approaches do not explicitly account for the uncertainty induced by unobserved modalities. As a consequence, the model may be trained with excessive confidence in the presence of missing modalities.

### 2.1 Theoretical Analysis of Probabilistic Representation

To explicitly model the uncertainty, we represent the task embedding $z_i$ in a probabilistic manner. The task representation $z_i$ is transformed into a Gaussian embedding composed of a mean $\mu_i = f_\theta(z_i)$ and a standard deviation $\sigma_i = f_\phi(z_i)$, where $f_\theta$ and $f_\phi$ are learnable networks parametrized by $\theta$ and $\phi$. The role of $\mu_i$ is to encode the task-specific embedding, while $\sigma_i$ models the uncertainty inherent in $\mu_i$. We sample $r_i$ from $\mathcal{N}(\mu_i, \sigma_i^2)$ to obtain a probabilistic representation that captures both the embedding and its uncertainty. To enable gradient optimization, we employ the reparameterization trick [13] on $r_i$ as

$$r_i = \mu_i + \epsilon \odot \sigma_i, \quad \epsilon \sim \mathcal{N}(0, I). \tag{1}$$

During training, task prediction is obtained as $\hat{y}_i = h(r_i)$ to incorporate uncertainty. At inference, only $\mu_i$ is used as $\hat{y}_i = h(\mu_i)$ for the consistent prediction.

Our training introduces stochastic perturbations around $\mu_i$, whose magnitude is governed by $\sigma_i$, thereby influencing the loss landscape. To clarify this effect, we

analyze how $\sigma_i$ affects the optimization of $\theta$, where $\mu_i{=}f_\theta(z_i)$, by examining the gradient of the loss $\mathcal{L}$ under sampling noise $\epsilon{\sim}\mathcal{N}(0, I)$ (i.e., $\nabla_\theta L(\theta;\epsilon)$). Applying the chain rule, gradient with respect to $\theta$ decomposes into $\nabla_{r_i} L(\theta;\epsilon)$ and $\frac{\partial r_i}{\partial \theta}$ as

$$\nabla_\theta L(\theta;\epsilon) = (\frac{\partial r_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \theta})^\top \nabla_{r_i} L(\theta;\epsilon) = \frac{\partial \mu_i}{\partial \theta}^\top \nabla_{r_i} L(\theta;\epsilon)\ (\because \frac{\partial r_i}{\partial \mu_i} = \mathbf{I}). \tag{2}$$

Since the noise $\epsilon{\odot}\sigma_i$ is added to $\mu_i$ in $r_i$, the gradient $\nabla_{r_i} L(\theta;\epsilon)$ can be linearized around the reference point $\mu_i$ (i.e., $\epsilon = 0$). A first-order Taylor expansion yields

$$\nabla_{r_i} L(\theta;\epsilon) \approx \nabla_{r_i} L(\theta;0) + \nabla^2_{r_i} L(\theta;0)(\epsilon \odot \sigma_i). \tag{3}$$

To characterize the influence of $\epsilon$ on $\nabla_\theta L(\theta;\epsilon)$, we first analyze the covariance of $\nabla_{r_i} L(\theta;\epsilon)$ with respect to the sampling noise $\epsilon$ (i.e., $\mathrm{Cov}_\epsilon[\nabla_{r_i} L(\theta;\epsilon)]$). Using the approximation in Eq. (3) and the covariance identity $\mathrm{Cov}[\mathbf{A}x]{=}\mathbf{A}\mathrm{Cov}[x]\mathbf{A}^\top$, the resulting gradient covariance with respect to $\epsilon$ is proportional to $\sigma_i^2$ as

$$\begin{aligned}\mathrm{Cov}_\epsilon[\nabla_{r_i} L(\theta;\epsilon)] &\approx \mathrm{Cov}_\epsilon[\nabla_{r_i} L(\theta;0)] + \mathrm{Cov}_\epsilon[\nabla^2_{r_i} L(\theta;0)(\epsilon \odot \sigma_i)] \\ &= \nabla^2_{r_i} L(\theta;0)\,\mathrm{diag}\big(\sigma_i^2\big)\nabla^2_{r_i} L(\theta;0)^\top\ (\because \mathrm{Cov}_\epsilon[\epsilon \odot \sigma_i] = \mathrm{diag}\big(\sigma_i^2\big)).\end{aligned} \tag{4}$$

The chain rule (Eq. (2)) further links $\mathrm{Cov}_\epsilon[\nabla_\theta L(\theta;\epsilon)]$ to $\mathrm{Cov}_\epsilon[\nabla_{r_i} L(\theta;\epsilon)]$ as

$$\mathrm{Cov}_\epsilon[\nabla_\theta L(\theta;\epsilon)] = \frac{\partial \mu_i}{\partial \theta}^\top \mathrm{Cov}_\epsilon[\nabla_{r_i} L(\theta;\epsilon)]\ \frac{\partial \mu_i}{\partial \theta}, \tag{5}$$

which also scales with $\sigma_i^2$ according to Eq. (4). Larger $\sigma_i$ induces higher gradient variance in the $\mu_i$-determining parameters $\theta$. The resulting stochastic updates fluctuate across iterations and partially cancel in expectation, which attenuate effective parameter updates and thereby diminish their impact on the task.

### 2.2 Set-Inclusive Modality Masking and Uncertainty Guidance

While $\sigma_i$ modulates the updates of the model parameter $\theta$, optimizing the task loss alone (i.e, $\mathcal{L} = \mathcal{L}_{\text{Task}}$) collapses $\sigma_i$ toward zero. To be explicit, analogous to the local expansion in Eq. (3), the expected loss admits the approximation as

$$\mathbb{E}_\epsilon[L(\mu_i + \epsilon \odot \sigma_i)]\ \approx\ L(\mu_i) + \tfrac{1}{2}\,\mathrm{Tr}\big(\nabla^2_{r_i} L(\theta;0)\,\mathrm{diag}(\sigma_i^2)\big)\,. \tag{6}$$

Since the second term is non-negative and strictly increasing in $\sigma_i^2$, minimizing the task loss drives $\sigma_i \to 0$, which necessitates additional regularization to preserve stochasticity in $r_i$. In this regime, we posit that $\sigma_i$ *should increase when task-relevant information is insufficient*, which naturally occurs *as available modalities are progressively removed.* This motivates our subsequent design, which incorporates set-inclusive training and uncertainty regularization.

**Set-Inclusive Modality Masking.** Let $\mathcal{M} = \{1, \ldots, M\}$ denote the index set of all modalities. For a given sample $x_i$, each modality configuration is represented as a subset $\mathcal{S}_i \subseteq \mathcal{M}$ indicating the observed modalities. The masking

indicator $\delta_i^{(m)}$ is set to 1 if and only if $m \in \mathcal{S}_i$, and to 0 otherwise. To enable set-inclusive uncertainty modeling, we consider three modality index sets associated with the same sample $x_i$. Let $\mathcal{S}_i^{\text{full}} = \mathcal{M}$ denote the full-modality configuration. We further define two partial modality sets $\mathcal{S}_i^k \in \{\mathcal{S}_i^{\text{sub}}, \mathcal{S}_i^{\text{sup}}\}$, which satisfy the inclusion relation as $\mathcal{S}_i^{\text{sub}} \subset \mathcal{S}_i^{\text{sup}} \subset \mathcal{S}_i^{\text{full}}$. For each partial modality set $\mathcal{S}_i^k$, the task representation $z_i^k$ is parameterized as a Gaussian embedding with mean $\mu_i^k$ and variance $(\sigma_i^k)^2$. We then perform stochastic sampling (Eq. (1)) as $r_i^k = \mu_i^k + \epsilon \odot \sigma_i^k$ to obtain the task prediction $\hat{y}_i^k = h(r_i^k)$. For a full-modality set, which serves as an uncertainty-free anchor, the prediction $\hat{y}_i^{\text{full}}$ is obtained directly from $\mu_i^{\text{full}}$. Accordingly, the task loss is defined over the modality sets as

$$\mathcal{L}_{\text{Task}} = \ell(\hat{y}_i^{\text{sub}}, y_i) + \ell(\hat{y}_i^{\text{sup}}, y_i) + \ell(\hat{y}_i^{\text{full}}, y_i), \tag{7}$$

where $\ell$ is a task-specific loss (e.g., Dice loss) and $y_i$ is the ground-truth label.

**Set-Inclusive Uncertainty Guidance.** As the full-modality configuration $S_i^{\text{full}}$ provides the most complete evidence, we treat its mean $\mu_i^{\text{full}}$ as an uncertainty-free anchor. For each incomplete configuration $k \in \{\text{sub}, \text{sup}\}$, we align its mean $\mu_i^k$ with this anchor, while allowing $\sigma_i^k$ to quantify the discrepancy between $\mu_i^k$ and $\mu_i^{\text{full}}$. We define the corresponding uncertainty-aware alignment loss $\mathcal{L}_{\text{UA}}$ as a Gaussian negative log-likelihood with gradients detached from $\mu_i^{\text{full}}$. For $\mu_i^k, \sigma_i^k \in \mathbb{R}^{C\times H\times W\times D}$ with channel dimension $C$ and spatial shape $H\times W\times D$, $\mathcal{L}_{\text{UA}}$ is computed element-wise and averaged over all elements via $\mathbb{E}$ as

$$\mathcal{L}_{\text{UA}} = \mathbb{E}\left[\sum_{k\in\{\text{sub},\text{sup}\}} \left(\frac{1}{2}\frac{(\mu_i^{\text{k}} - \text{GradStop}(\mu_i^{\text{full}}))^2}{(\sigma_i^{\text{k}})^2} + \frac{1}{2}\log(\sigma_i^{\text{k}})^2\right)\right]. \tag{8}$$

The former term penalizes deviations of $\mu_i^k$ from $\mu_i^{\text{full}}$, while enlarging $\sigma_i^k$ when the discrepancy is substantial. The latter prevents $\sigma_i^k$ from growing unbounded. These terms drive $\mu_i^k$ toward $\mu_i^{\text{full}}$ while guiding $\sigma_i^k$ to reflect their discrepancy. We further impose an ordering constraint consistent with the set-inclusion hierarchy. Since $\mathcal{S}_i^{\text{sub}} \subset \mathcal{S}_i^{\text{sup}}$, uncertainty from $\mathcal{S}_i^{\text{sub}}$ should be greater. To penalize violations of this hierarchy, we introduce an uncertainty ordering loss $\mathcal{L}_{\text{UO}}$ as

$$\mathcal{L}_{\text{UO}} = \mathbb{E}\left[\max\big(0, (\sigma_i^{\text{sup}})^2 - (\sigma_i^{\text{sub}})^2\big)\right]. \tag{9}$$

The overall objective $\mathcal{L}$ combines the task loss with our two uncertainty regularization terms, weighted by hyperparameters $\alpha$ and $\beta$, respectively, as

$$\mathcal{L} = \mathcal{L}_{\text{Task}} + \alpha\mathcal{L}_{\text{UA}} + \beta\mathcal{L}_{\text{UO}}. \tag{10}$$

## 3 Experimental Settings

### 3.1 Datasets and Evaluation Metrics

We evaluate our framework on two brain tumor segmentation benchmarks, BraTS 2018 and 2020 [17], to assess performance across cohorts of different scales and compositions. Both datasets provide four MRI modalities (i.e., T1, T1c, T2,

**Table 1.** Test performance measured by the Dice similarity coefficient (DSC, %) on the BraTS 2020 dataset. For each modality, • and ○ denote its presence and absence, respectively. The best and second-best results are highlighted in **bold** and underline.

| Modalities | | | | Whole tumor (WT) | | | | | Tumor core (TC) | | | | | Enhancing tumor (ET) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| F | T1 | T1c | T2 | RFNet | mmFormer | M$^3$AE | DC-Seg | Ours | RFNet | mmFormer | M$^3$AE | DC-Seg | Ours | RFNet | mmFormer | M$^3$AE | DC-Seg | Ours |
| ○ | ○ | ○ | • | 86.05 | 85.51 | 86.10 | 86.72 | **87.24** | 71.02 | 63.36 | 71.80 | 70.88 | **73.23** | 46.29 | 49.09 | 47.10 | 47.76 | **49.55** |
| ○ | ○ | • | ○ | 76.77 | 78.04 | 78.90 | 79.54 | **80.11** | 81.51 | 81.51 | 83.60 | 84.62 | **86.15** | 74.85 | 78.30 | 73.60 | 78.90 | **81.93** |
| ○ | • | ○ | ○ | 77.16 | 76.24 | 79.00 | 78.47 | **80.19** | 66.02 | 63.23 | **69.40** | 66.63 | 64.88 | 37.30 | 37.62 | 40.40 | **42.19** | 41.63 |
| • | ○ | ○ | ○ | 87.32 | 86.54 | **88.00** | 87.80 | 87.85 | 69.19 | 64.60 | 68.70 | 71.27 | **72.07** | 38.15 | 36.68 | 40.20 | **41.66** | 40.56 |
| ○ | ○ | • | • | 87.74 | 87.52 | 87.10 | 88.17 | **88.43** | 83.45 | 82.69 | 85.60 | 86.34 | **86.89** | 75.93 | 77.20 | 76.00 | 80.43 | **83.31** |
| ○ | • | • | ○ | 81.12 | 80.70 | 80.10 | 82.22 | **83.41** | 83.40 | 82.81 | 83.80 | 85.18 | **87.02** | 78.01 | 81.71 | 75.30 | 79.25 | **83.11** |
| • | • | ○ | ○ | 89.73 | 88.76 | 89.60 | 90.01 | **90.20** | 73.07 | 71.76 | 72.80 | 74.50 | **74.71** | 40.98 | 42.98 | 43.70 | **46.90** | 46.81 |
| ○ | • | ○ | • | 87.73 | 86.94 | 87.30 | 88.09 | **88.67** | 73.13 | 67.76 | 72.90 | 73.09 | **73.53** | 45.65 | 49.12 | 48.70 | **50.19** | 48.86 |
| • | ○ | ○ | • | 89.87 | 89.49 | 90.10 | 90.32 | **90.42** | 74.14 | 70.34 | 74.30 | 75.11 | **76.34** | 49.32 | 49.06 | 47.10 | **51.32** | 50.82 |
| • | ○ | • | ○ | 89.89 | 89.31 | 89.50 | 89.99 | **90.44** | 84.65 | 83.79 | 85.50 | 85.90 | **86.44** | 76.67 | 79.44 | 75.90 | 80.28 | **82.25** |
| • | • | • | ○ | 90.69 | 89.79 | 89.60 | 90.65 | **91.09** | 85.07 | 84.44 | 85.60 | 86.29 | **86.40** | 76.81 | 80.65 | 76.30 | 81.41 | **83.20** |
| • | • | ○ | • | 90.60 | 89.83 | 90.20 | 90.77 | **91.00** | 75.19 | 72.42 | 74.40 | 75.53 | **75.98** | 49.92 | 50.08 | 48.20 | **52.05** | 50.80 |
| • | ○ | • | • | 90.68 | 90.49 | 90.50 | 90.62 | **91.27** | 84.97 | 83.94 | 85.80 | 86.21 | **86.39** | 77.12 | 78.73 | 77.40 | 79.42 | **81.54** |
| ○ | • | • | • | 88.25 | 87.64 | 87.40 | 88.73 | **89.14** | 83.47 | 83.66 | 85.80 | 86.49 | **86.96** | 76.99 | 77.34 | 78.00 | 81.66 | **83.55** |
| • | • | • | • | 91.11 | 90.54 | 90.40 | 90.95 | **91.58** | 85.21 | 84.61 | 86.20 | **86.46** | 86.38 | 78.00 | 79.92 | 77.50 | 81.52 | **83.35** |
| Average | | | | 86.98 | 86.49 | 86.90 | 87.54 | **88.07** | 78.23 | 76.06 | 79.10 | 79.63 | **80.22** | 61.47 | 63.19 | 61.70 | 65.00 | **66.02** |

**Table 2.** Average Dice similarity coefficient (DSC, %) on BraTS 2018 dataset.

| Methods | WT | TC | ET |
|---|---|---|---|
| RFNet | 85.49 | 76.00 | 58.44 |
| mmFormer* | 86.20 | 73.65 | 56.20 |
| M$^3$AE | 85.82 | 77.37 | 59.85 |
| DC-Seg* | 86.59 | 77.14 | 59.55 |
| Ours | **87.10** | **78.37** | **62.52** |

**Table 3.** Ablation study on BraTS 2020. (*Set&Prob*: set-inclusive probabilistic design)

| *Set&Prob* | $\mathcal{L}_{\text{UA}}$ | $\mathcal{L}_{\text{UO}}$ | WT | TC | ET |
|---|---|---|---|---|---|
| × | × | × | 86.98 | 78.23 | 61.47 |
| ✓ | × | × | 87.54 | 79.20 | 63.74 |
| ✓ | ✓ | × | 87.73 | 79.32 | 65.57 |
| ✓ | × | ✓ | 87.57 | 79.31 | 64.61 |
| ✓ | ✓ | ✓ | **88.07** | **80.22** | **66.02** |

and FLAIR) and standardized tumor annotations, and performance is evaluated on WT, TC, and ET. All volumes are skull-stripped, aligned to a common anatomical template, and resampled to an isotropic resolution of 1mm$^3$, followed by zero-mean, unit-variance normalization within the brain region. BraTS 2018 contains 285 subjects, whereas BraTS 2020 extends the cohort to 369 subjects.

We follow the same train, validation, and test splits as M$^3$AE [15] for BraTS 2018 (199:29:57) and RFNet [8] for BraTS 2020 (219:50:100). During training, volumetric patches of 112×112×112 are randomly extracted. At inference, we adopt the sliding-window strategy of RFNet [8] and average overlapping predictions. Performance is evaluated using the Dice similarity coefficient (DSC) [16].

### 3.2 Baselines and Implementation Details

To validate our work, we compare our method with recent state-of-the-art approaches [8,14,15,21]. We report the results on each dataset from the original papers when available. Otherwise, we reproduce them using the official code and hyperparameters. These reproduced results are marked with * in Tab. 2.

Following the baselines [14,15,21], we adopt the backbone architecture and training protocol of RFNet [8]. The output of the penultimate layer (i.e., the feature representation before the segmentation head) is replaced with a probabilistic embedding. Two separate 3D CNN layers are used for $f_\theta$ and $f_\phi$, respectively.

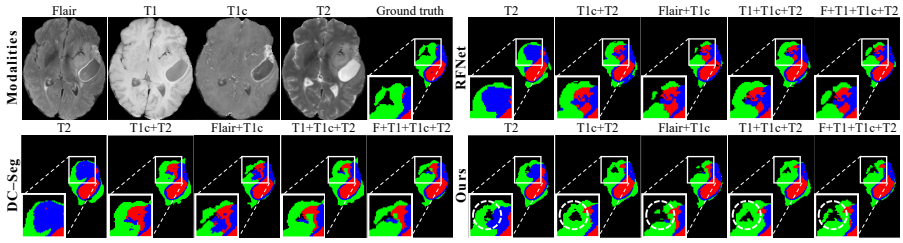


**Fig. 2.** Qualitative results on BraTS 2020. Our model produces improved segmentations over RFNet [8] and DC-Seg [14] across various configurations. (Region: ET (blue) ⊂ TC (blue, red) ⊂ WT (blue, red, green), Sample: *"HG_ BraTS20_ Training_ 357"*)

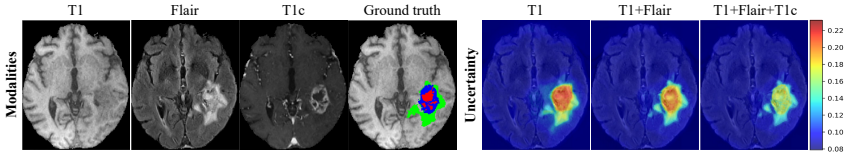


**Fig. 3.** Visualization of input modalities and ground-truth mask, with trained uncertainty overlaid on T1. Uncertainty, initially higher within tumor regions, decreases as additional modalities are incorporated. (Sample: *"HG_ BraTS20_ Training_ 045"*)

Our framework was trained for 800 epochs using the Adam optimizer with a learning rate of $2\times10^{-4}$ and a batch size of 2, where $\alpha$=0.01 and $\beta$=5.

## 4 Results and Analysis

### 4.1 Performance on the BraTS 2018 and 2020 datasets

**Quantitative Results.** We first compare our method with recent baselines on BraTS 2020 dataset in Tab. 1. Our method outperformed the baselines by a clear margin in terms of the averaged Dice similarity score across the three tumor subregions. In particular, it consistently achieved the best or second-best performance under almost every modality configuration, which demonstrates its robustness against diverse missing-modality scenarios. A similar trend was observed on BraTS 2018 dataset, as shown in Tab. 2. These consistent improvements across the datasets indicate our superiority across underlying data cohorts.
**Qualitative Results.** We visualize center-slice predictions on BraTS 2020 under modality configurations, together with RFNet [8] and DC-Seg [14] (Fig. 2). Our model yields more accurate and consistent segmentations across various configurations, notably preserving the background regions adjacent to the whole tumor. In contrast, the baselines overfill these regions (in blue) and overestimate the ET. These results indicate that our model relies on reliable cues, whereas deterministic embeddings in the baselines yield spurious hallucinations.

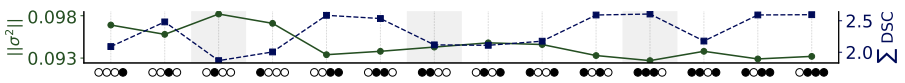


**Fig. 4.** Variance magnitude (—) and summed test DSC across tumor subregions (---) for each modality configuration on BraTS 2020. • and ◦ indicate observed and missing modalities (Flair, T1, T1c, T2). Higher uncertainty correlates with lower performance. The shaded regions highlight that a superset exhibits lower uncertainty than its subset.

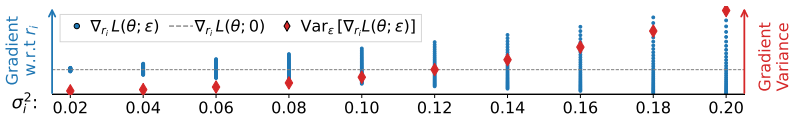


**Fig. 5.** Effect of variance magnitude on gradient behavior under controlled scaling.

### 4.2 Model Behavior Analyses

**Ablation Study.** We analyze the contribution of each component over the deterministic baseline (i.e., RFNet [8]) on BraTS 2020 (Tab. 3). While adopting probabilistic embeddings with set-inclusive modality masking improves segmentation, the variance collapse restricts its benefit (see Eq.(6)). $\mathcal{L}_{\mathrm{UA}}$ mitigates this collapse and further enhances performance, whereas $\mathcal{L}_{\mathrm{UO}}$ alone brings marginal gains as it constrains ordering rather than magnitude. Combining $\mathcal{L}_{\mathrm{UA}}$ and $\mathcal{L}_{\mathrm{UO}}$ yields the best results by providing complementary uncertainty supervision.
**Uncertainty Analysis.** We first examine the spatial distribution of learned uncertainty by averaging $\sigma$ across channels and overlaying it on the T1 image (Fig. 3). Elevated uncertainty is primarily localized within tumor regions, particularly the enhancing tumor (blue label). By incorporating Flair and T1c together, uncertainty in these regions progressively decreases, indicating that complementary evidence reduces ambiguity and stabilizes model updates.

We further investigate the relationship between uncertainty and performance on BraTS 2020 across modality configurations (Fig. 4). Uncertainty is quantified by the mean squared magnitude of $\sigma$ during training, denoted as $||\sigma^2||$. Performance is measured by the summed Dice similarity coefficient over the three tumor subregions, denoted as $\sum$DSC. A clear inverse relationship is observed that configurations with larger variance magnitudes consistently correspond to lower test performance, and vice versa. This trend suggests that when missing modalities induce greater task-relevant information loss, the model responds by increasing uncertainty rather than committing to incomplete evidence.

Moreover, as highlighted by the shaded samples, superset configurations consistently exhibit lower uncertainty and higher performance than their subsets. This ordering behavior *aligns with our set-inclusive design*, confirming that uncertainty systematically reflects modality-induced information incompleteness.
**Gradient Scaling Analysis.** To empirically validate Eq. (4), which predicts that uncertainty $\sigma_i$ modulates gradient covariance, we analyze the $\nabla_{r_i}\mathcal{L}(\theta;\epsilon)$ at an arbitrary element of $r_i$ with fixed $\mu_i$ while varying $\sigma_i$ (Fig. 5). For each $\sigma_i$, we

sample $\epsilon$ 30 times to compute $\nabla_{r_i}\mathcal{L}(\theta;\epsilon)$. Compared to deterministic gradient ($\epsilon = 0$; dashed line), the sampled ones become increasingly dispersed as $\sigma_i^2$ grows (i.e., enlarged variance). This confirms the theoretical relationship in Eq. (4).

## 5 Conclusion

In this work, we address the intrinsic uncertainty induced by missing modalities through a probabilistic framework for incomplete multimodal brain tumor segmentation. This design explicitly reduces reliance on spurious cues, thereby yielding more reliable predictions. Extensive experiments on BraTS 2018 and 2020 demonstrate robustness across missing-modality scenarios. Both theoretical and empirical analyses support the necessity of our uncertainty modeling.

**Acknowledgments.** This research was supported by RS-2025-02216257 (60%), RS-2026-25494850 (30%), RS-2019-II1091906 (AI Graduate Program at POSTECH, 5%), and RS-2022-II220290 (5%).

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

## References

1. Baek, S., Sim, J., Dere, M., et al.: Modality-agnostic style transfer for holistic feature imputation. In: ISBI. pp. 1–5. IEEE (2024)
2. Baek, S., Sim, J., Wu, G., et al.: Ocl: Ordinal contrastive learning for imputating features with progressive labels. In: MICCAI. pp. 334–344. Springer (2024)
3. Baid, U., Ghodasara, S., Mohan, S., et al.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021)
4. Bakas, S., Sako, C., Akbari, H., et al.: The university of pennsylvania glioblastoma (upenn-gbm) cohort: advanced mri, clinical, genomics, & radiomics. Scientific data **9**(1), 453 (2022)
5. Bakas, S., Shukla, G., Akbari, H., et al.: Overall survival prediction in glioblastoma patients using structural magnetic resonance imaging (mri): advanced radiomic features may compensate for lack of advanced mri modalities. JMI **7**(3), 031505–031505 (2020)
6. Blystad, I., Warntjes, J.M., Smedby, Ö., et al.: Quantitative mri for analysis of peritumoral edema in malignant gliomas. PLoS One **12**(5), e0177135 (2017)
7. Conte, G.M., Weston, A.D., Vogelsang, D.C., et al.: Generative adversarial networks to synthesize missing t1 and flair mri sequences for use in a multisequence brain tumor segmentation model. Radiology **299**(2), 313–323 (2021)
8. Ding, Y., Yu, X., Yang, Y.: Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In: ICCV. pp. 3975–3984 (2021)
9. Dorent, R., Joutard, S., Modat, M., et al.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: MICCAI. pp. 74–82. Springer (2019)

10. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Medical image analysis **35**, 18–31 (2017)
11. Işın, A., Direkoğlu, C., Şah, M.: Review of mri-based brain tumor image segmentation using deep learning methods. Procedia computer science **102**, 317–324 (2016)
12. Kamnitsas, K., Ledig, C., Newcombe, V.F., et al.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017)
13. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
14. Li, H., Li, Z., Mao, Y., et al.: Dc-seg: Disentangled contrastive learning for brain tumor segmentation with missing modalities. In: MICCAI. pp. 138–148. Springer (2025)
15. Liu, H., Wei, D., Lu, D., et al.: M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In: AAAI. vol. 37, pp. 1657–1665 (2023)
16. Lr, D.: Measures of the amount of ecologic association between species. Ecology **26**, 297–302 (1945)
17. Menze, B.H., Jakab, d., Bauer, S., et al.: The multimodal brain tumor image segmentation benchmark (brats). TMI **34**(10), 1993–2024 (2014)
18. Pereira, S., Pinto, A., Alves, V., et al.: Brain tumor segmentation using convolutional neural networks in mri images. TMI **35**(5), 1240–1251 (2016)
19. Sharma, A., Hamarneh, G.: Missing mri pulse sequence synthesis using multi-modal generative adversarial network. TMI **39**(4), 1170–1183 (2019)
20. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. NeurIPS **30** (2017)
21. Zhang, Y., He, N., Yang, J., et al.: mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In: MICCAI. pp. 107–117. Springer (2022)