# Tackling Dual-stage Missing Modalities in Brain Tumor Segmentation via Robust Modality Reconstruction and Prompt-guided Modality Adaptation

**Yunpeng Zhao[1*], Cheng Chen[3,4*], Qing You Pang[5], Yibing Fu[1], Quanzheng Li[7], Carol Tang[5,6], Beng Ti Ang[5,6], Yueming Jin[1,2†]**

[1]Department of Biomedical Engineering, National University of Singapore
[2]Department of Electrical and Computer Engineering, National University of Singapore
[3]Department of Electrical and Electronic Engineering, The University of Hong Kong
[4]School of Biomedical Engineering, The University of Hong Kong
[5]National Neuroscience Institute
[6]Duke-National University of Singapore Medical School
[7]Massachusetts General Hospital, Harvard Medical School
{yunpeng.zhao,yibingfu,carol.tang}@u.nus.edu, cchen@eee.hku.hk, eddy_pq_you@nni.com.sg,
li.quanzheng@mgh.harvard.edu, ang.beng.ti@singhealth.com.sg, ymjin@nus.edu.sg

## Abstract

Addressing missing modalities is a critical challenge in multimodal brain tumor segmentation. Most existing approaches merely handle modality-incomplete inputs during inference, assuming a full set of modalities for all training samples. However, this unrealistic assumption limits the usage of abundant modality-incomplete data commonly observed in clinical practice. In this paper, we explore a more practical task of tackling missing modalities during both training and inference. We propose a universal model featuring robust modality reconstruction and prompt-guided modality adaptation. Our mask-reconstruction pre-training enables robust modality-invariant representation learning, during which we design a novel distribution approximation method that supervises the reconstruction of absent modalities without requiring full-modal training data. Afterwards, when adapting our model to the segmentation task, we introduce the complete-then-distill (CTD) paradigm, which first estimates missing modalities in training samples from the available ones, and then distills the knowledge from the reconstructed full-modal representations to enhance learning from modality-incomplete data. Moreover, we propose prompt-guided modality adaptation to personalize a subset of model parameters during CTD, enabling the model to adapt to each distinct modality input scenario by using prompts with rich visual-textual information. Extensive experiments on two brain tumor segmentation benchmarks show our method consistently surpasses previous state-of-the-art approaches under dual-stage missing modality settings across various missing ratios.

## Introduction

Brain tumor segmentation, aiming to identify tumor regions, is clinically crucial for diagnosis, progress monitoring, and surgical planning (Pereira et al. 2016; Ranjbarzadeh et al. 2021). Different magnetic resonance imaging (MRI) modalities, including T1-weighted (T1), contrast-enhanced T1 (T1ce), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR), are commonly adopted to provide complementary information for accurate segmentation (Liu et al. 2022; Xing et al. 2022; Isensee et al. 2021; Zhang et al. 2020). However, in real-world clinical scenarios, the occurrence of one or more missing modalities is prevalent due to artifacts, allergies to contrast agents, or economic considerations (Konwer et al. 2023; Liu et al. 2023a). Missing modalities can significantly hinder both the training and inference processes of conventional multimodal learning methods, as these methods are unable to leverage modality-incomplete data during training and struggle to maintain robust performance when modalities are missing at inference time. Hence, handling missing modalities effectively at both the **training** and **inference** stages is crucial for robust multimodal brain tumor segmentation.

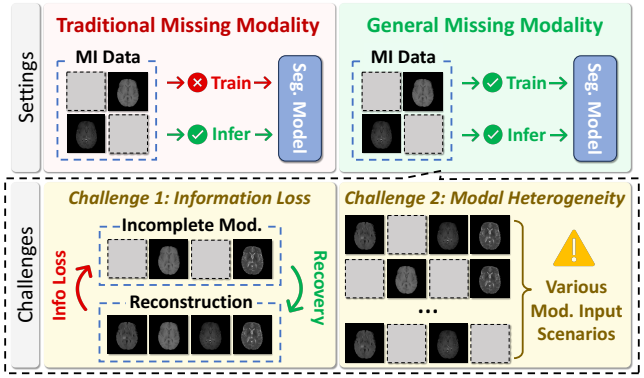


Figure 1: Comparison between the traditional inference-stage missing modality problem and the general dual-stage setting. "MI" denotes "Modality-Incomplete".

As illustrated in Fig. 1, most of the existing approaches addressing missing modalities merely focus on modality-

*These authors contributed equally.
†Corresponding author.

incomplete data during inference (Liu et al. 2023a; Zhang et al. 2022; Ding, Yu, and Yang 2021; Wang et al. 2021). These methods typically assume that all training samples have a complete set of modalities, and are not optimized to fully leverage the modality-incomplete data for model training. Although many existing methods can be adapted to accommodate modality-incomplete training data through zero-padding, this strategy fails to recover the missing information and thus limits the model's ability to fully exploit the potential of multimodal learning. Moreover, these methods often rely on full-modal regularization to learn an unbiased common latent space. When using a heterogeneous training set composed of both modality-complete and incomplete samples, the learned representations could be biased.

Currently, research on handling modality-incomplete data for segmentation during both training and inference remains very limited. Konwer et al. (2023) adopt meta-learning to treat each modality combination as a separate task, using limited full-modal samples to enhance partial-modal representations, but it relies on having enough examples for each missing modality combination to form meaningful meta-tasks, limiting its effectiveness to a certain range of missing modality ratios. M$^3$FeCon (Zeng et al. 2024) always treats missing modalities in training data as masked in the masked autoencoder, reconstructing all modalities at the feature level for segmentation. However, it lacks supervision for reconstructing missing modalities, often resulting in low-quality, inaccurate feature estimates and causing biased representation learning due to modality imbalance.

Therefore, it is a critical yet challenging task to develop a universal model that seamlessly supports both modality-complete and incomplete data throughout the training and inference stages. To address this challenge, we identify two key obstacles. **First**, the loss of critical information due to missing modalities is an obvious issue, which has received considerable attention in prior research. However, our challenge extends beyond just the testing phase, as information loss also occurs during training. This dual-phase information loss imposes more stringent demands and challenges on our method's design. **Second**, the varying missing modality combinations during training and testing result in highly heterogeneous data distributions, a factor often overlooked by previous works. For example, in brain tumor segmentation, four modalities can result in fifteen possible missing modality combinations, each corresponding to a distinct data distribution. Prior studies in multimodal learning have shown that forcing a model to handle such diverse distributions using fully shared parameters is suboptimal (Dou et al. 2020).

In this paper, we propose *a universal model with robust modality reconstruction and prompt-based modality adaptation*, to address the two critical challenges. For information loss, we first pre-train our model by the mask-reconstruction paradigm, which can learn modality-invariant representations and exploit inter-modal correlations to estimate missing information. Given the absence of supervision for reconstructing missing modalities during training, we innovate distribution approximation, which adopts segmentation loss to guide the distribution of the reconstructed modalities to approximate that of real full modalities. Then when adapting the pre-trained model to segmentation tasks, we design a complete-then-distill paradigm to effectively utilize the reconstructed full-modal information to enhance model learning with missing modalities and ease the modality-biased problem mentioned by Konwer et al. (2023).

To further address data distribution heterogeneity arising from various missing modality combinations, we enable the model to adapt a specific subset of parameters optimized for each missing-modality scenario. Specifically, we propose an innovative prompt-guided modality adaptation method, which encodes the information of available modalities via combined visual features and textual prompts as input indicators. The indicators guide the hyper-network to adaptively generate scenario-specific parameters, allowing the model to effectively handle arbitrary modality combinations.

Overall, our contributions can be summarized as follows:

- We introduce a universal model to address dual-stage missing modalities, which adequately utilizes modality-incomplete training data and enables robust inference with incomplete testing inputs.
- We propose distribution approximation to provide supervision for the reconstruction of absent modalities during modality-invariant representation learning and a complete-then-distill strategy to fully leverage the estimated full-modal information when adapting our model to the segmentation task.
- To address data heterogeneity, we design a prompt-guided modality adaptation mechanism that can personalize a subset of model parameters for each unique missing modality scenario based on visual-textual prompts indicating modality combination information.

## Related Work

### Missing Modalities in Medical Image Segmentation

**Inference-stage missing modalities.** Knowledge distillation (KD) approaches (Hu et al. 2020; Wei, Luo, and Luo 2023; Wang et al. 2021; Azad, Khosravi, and Merhof 2022) transfer knowledge from teacher models trained on full-modal data to student models with missing modalities. Generative methods (Yang, Sun, and Xu 2023; Lee, Moon, and Ye 2020; Sharma and Hamarneh 2019) synthesize missing modalities using available ones to perform segmentation with the completed data. Shared latent space models (Ding, Yu, and Yang 2021; Zhang et al. 2022; Chen et al. 2019a; Zhou et al. 2021; Liu et al. 2023a) encode all modalities into a common latent subspace to learn modal-invariant representations. However, most existing methods lack specific designs for training-stage missing modalities. KD and generative approaches require full-modal data to train the teacher or generator. Shared latent space models also strongly depend on complete training data to regularize the common latent space (Konwer et al. 2023). Although much attention has been paid to missing modalities at test time, little focus has been given to training-stage modality incompleteness.

**Training-stage missing modalities.** Konwer et al. (2023) incorporated both modality-complete and incomplete train-

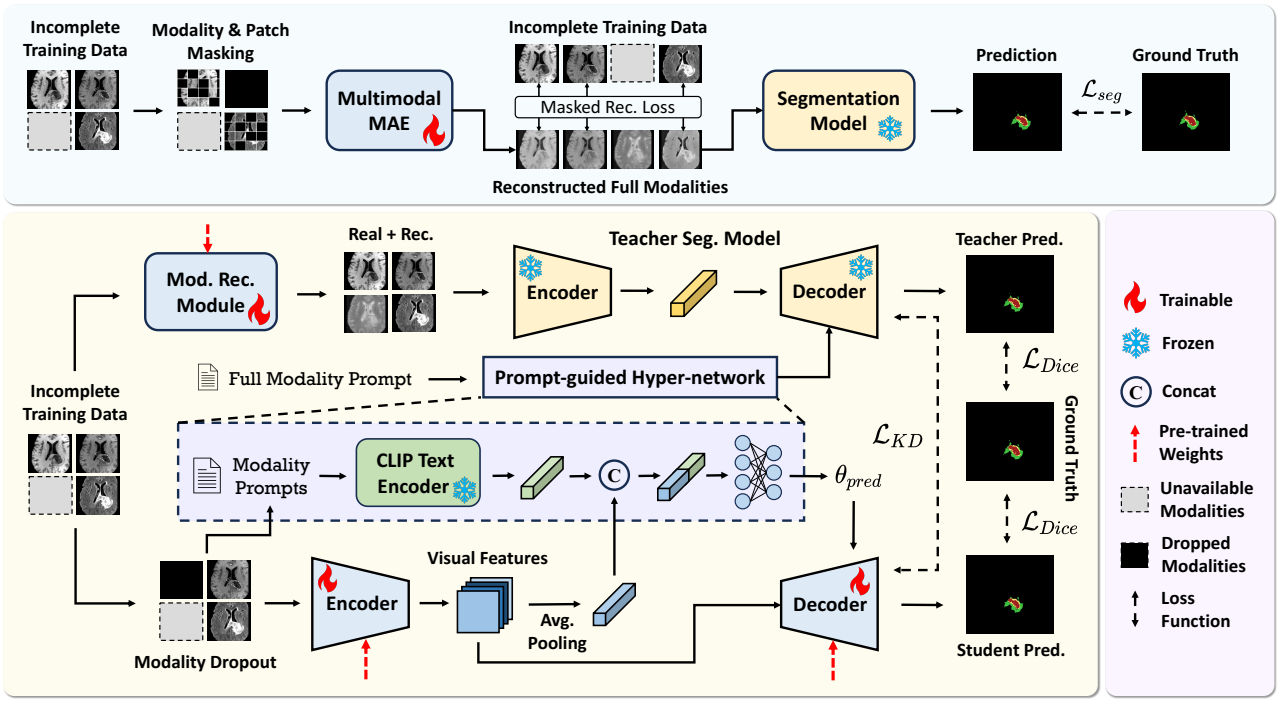


Figure 2: Overview of our proposed universal model for tackling missing modalities during training and inference.

ing data through a meta-learning framework, which can effectively use full-modal samples to enhance partial-modal training data. Yet, it lacks flexibility and its performance degrades substantially with insufficient full-modal data for meta-test. M$^3$FeCon (Zeng et al. 2024) applies a mask-reconstruction paradigm to modality features, aiming to recover full-modal features from randomly masked partial-modal ones, where the absent modalities in the training data are always treated as masked. The feature reconstruction is only supervised on the available training modalities and the reconstructed full-modal features are then used to perform segmentation. However, simply ignoring the reconstruction supervision for absent training modalities will lead to low-quality feature estimates, which directly results in suboptimal performance when full-modal samples are scarce.

### Modality Adaptation via Model Personalization

Due to high modality heterogeneity in missing modality scenarios, sharing the same model parameters across different modality combinations often leads to suboptimal results (Dou et al. 2020). Recent efforts have explored adaptive model personalization. Flex-MoE (Yun et al. 2024) used sparse mixture-of-experts for disease diagnosis, activating a specific set of experts for each missing modality scenario. But it required sufficient full-modal data for training, making it unsuitable for dual-stage missing modality settings. Hyper-GAE (Yang, Sun, and Xu 2023) introduced a hypernetwork design to adjust model parameters based on binary indicators representing modality availability. However, it treated modalities as entirely independent and failed to semantically model the collaboration between them.

## Methodology

In this section, we introduce our framework for dual-stage missing modalities in multimodal brain tumor segmentation, as shown in Fig. 2. In pre-training, our model reconstructs all four modalities from masked inputs, applying reconstruction loss only to available training modalities, while our distribution approximation supervises the reconstruction of missing modalities at the distribution level. The pre-trained model not only provides initialization for the segmentation task as it has learned modality-invariant representations, but also naturally serves as a modality recovery module. When adapting our model to segmentation tasks, we propose complete-then-distill paradigm. The student model is trained with modality dropout, while the teacher receives full-modal data completed by the modality recovery module; the teacher's full-modal representations enhance the student's partial-modal ones, and the modality recovery module is continually refined during training. Prompt-guided modality adaptation is integrated into both teacher and student models, enabling flexible adaptation to various missing modality scenarios and addressing data heterogeneity. During inference, we only keep the student model.

### Distribution Approximation in Pretraining

A key challenge in addressing dual-stage missing modalities is how to enable the model to learn modality-robust representations that benefit modality-incomplete inference and recover absent modality information in the training data. We have identified mask-reconstruction paradigm as an exceptionally suitable solution, as it not only learns modality-invariant features through pre-training but also naturally re-

constructs full-modal data during this process. However, for modality-incomplete training data, missing modalities lack ground truth for reconstruction supervision. Simply ignoring them and calculating the loss only on modalities with ground truths leads to poor reconstruction quality for the missing modalities in the training data. Furthermore, modality-imbalanced supervision during pre-training results in suboptimal learning of modality-invariant representations, which affects the model's robustness to missing modalities during inference. Formally, we denote the train set $\mathcal{D} = \{\mathcal{D}_f, \mathcal{D}_m\}$, where $\mathcal{D}_f$ and $\mathcal{D}_m$ denotes the subset of full-modal and modality-incomplete data respectively. Let $M$ represent the total number of modalities. For a training sample $\boldsymbol{x} \in \mathcal{D}_m$, where $\boldsymbol{x} = \{x_i\}_{i=1}^{M}$ and $x_i = \mathbf{0}$ if the $i$-th modality is missing, the reconstruction loss on the available modalities is:

$$\mathcal{L}_{rec}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{|\{x_i | x_i \neq \mathbf{0}\}|} \sum_{i \in \{i | x_i \neq \mathbf{0}\}} \|\hat{x}_i - x_i\|^2, \quad (1)$$

where $\hat{\boldsymbol{x}}$ is the reconstructed full modalities.

To regularize the reconstruction of missing modalities during training, we propose distribution approximation that employs a pre-trained multimodal segmentation model[1] and uses the segmentation loss to provide modality distribution-level supervision without relying on modality reconstruction ground truth. The underlying assumption is that a multimodal segmentation model with fair generalization should exhibit superior performance (i.e., the lowest loss) when provided with full-modal input, compared to any other partial-modal inputs. In this regard, if we fix the segmentation model's parameters and minimize the segmentation loss to optimize the modality reconstruction process, we can maximize the probability that the reconstructed modalities conform to the real full-modal distribution. Through distribution approximation, we shift the supervision of reconstruction from relying on modality ground truth to relying on segmentation labels, which are available even in modality-incomplete training data. This can be formulated as:

$$\min_{\hat{\boldsymbol{x}}} \mathcal{L}_{seg}(f(\hat{\boldsymbol{x}}), \boldsymbol{y}), \quad (2)$$

where $\boldsymbol{y}$ is the segmentation label, $f(\cdot)$ denotes the pre-trained segmentation model, and $\mathcal{L}_{seg}$ is segmentation loss, e.g., cross-entropy. Then, our mask-reconstruction pretraining with distribution approximation can be represented as:

$$\min_{\theta} \mathcal{L}_{rec}(\phi_\theta(\texttt{MASK}(\boldsymbol{x})), \boldsymbol{x}) + \lambda \mathcal{L}_{seg}(f(\phi_\theta(\texttt{MASK}(\boldsymbol{x}))), \boldsymbol{y}), \quad (3)$$

where $\phi_\theta(\cdot)$ denotes our model during pre-training parameterized by $\theta$, $\texttt{MASK}(\cdot)$ is the masking operation, and $\lambda$ is a factor set to 0.1 that controls the weight of supervision provided by distribution approximation.

## Complete-then-Distill Paradigm

After mask-reconstruction pre-training, we adapt the pre-trained model, which has learned robust modality-invariant representations, on the segmentation task. Following the common practice in missing modality research, we apply modality dropout during training to help the model adapt to modality-incomplete inputs during testing. For missing modalities in the training data, a straightforward solution is using the pre-trained model as a modality recovery module to complete absent modalities. However, we observe discrepancies between the quality of the reconstructed missing modalities and that of those available real modalities (see Appendix). Directly combining reconstructed missing modalities and available modalities as full-modal samples for training leads to suboptimal performance.

To address this problem, we propose the complete-then-distill paradigm. Previous studies show that imperfect synthetic data can be used to distill the teacher model and improve the student model's generalization (Mordvintsev, Olah, and Tyka 2015; Chen et al. 2019b). Inspired by this, given that the full-modal data completed by modality recovery module consists of high-quality real modalities and relatively low-quality reconstructed modalities, we adopt knowledge distillation to effectively utilize the recovered full-modal information to compensate for information loss caused by modality-incomplete training data.

As shown in Fig. 2, our complete-then-distill paradigm consists of a modality recovery module $\mathcal{M}_{\text{rec}}$, a frozen teacher model $\mathcal{T}$, and a learnable student segmentation model $\mathcal{S}$. The $\mathcal{M}_{\text{rec}}$ and $\mathcal{S}$ are both initialized by pre-trained parameters. We first use $\mathcal{M}_{\text{rec}}$ to estimate the absent modalities $\hat{x}_m$ and combine them with available real modalities $x_a$ to recover full-modal data $\hat{x}_{\text{full}}$. The $\hat{x}_{\text{full}}$ is then fed into $\mathcal{T}$ to extract multi-scale features $\mathbf{F}_l^{\mathcal{T}} = \mathcal{T}_l(\hat{x}_{\text{full}}),\ l = 1, 2, \ldots, L$. The student model's input during training is obtained by applying modality dropout to $x_a$. Therefore, the multi-scale features in student model $\mathbf{F}_l^{\mathcal{S}} = \mathcal{S}_l(\texttt{MDrop}(x_a)),\ \ l = 1, 2, \ldots, L$, where $\texttt{MDrop}(\cdot)$ refers to modality dropout. We perform knowledge distillation on the multi-scale features, leveraging the recovered full-modal data to activate the full-modal knowledge within the teacher model, thereby enhancing the student model's partial-modal features:

$$\mathcal{L}_{\text{KD}} = \sum_{l=1}^{L} ||\mathbf{F}_l^{\mathcal{T}} - \mathbf{F}_l^{\mathcal{S}}||_2^2. \quad (4)$$

To further refine the missing modality estimation quality, $\mathcal{M}_{\text{rec}}$ is learnable during the complete-then-distill process and is continually tuned via distribution approximation:

$$\mathcal{L}_{\mathcal{M}_{\text{rec}}} = \mathcal{L}_{seg}(\mathcal{T}(\hat{x}_{\text{full}}), y), \quad (5)$$

where $\mathcal{L}_{seg}$ here is a mixture of Dice loss and cross-entropy loss, and $y$ is the segmentation label.

## Prompt-guided Modality Adaptation

Beyond information loss, another challenge posed by missing modalities is the significant heterogeneity in modality distributions. We propose prompt-guided modality adaptation as a solution. It adopts a hyper-network to adaptively adjust a subset of model parameters for different missing modality scenarios according to visual-textual modality prompts. Hyper-network (Ha, Dai, and Le 2017) involves

[1] In practice, the segmentation model can be obtained by training a model on $\mathcal{D}$ (this work) or using publicly available models.

using a simple neural network to generate parameters for another primary network, which allows the primary network to adapt its weights on the fly to different tasks or data points.

Concretely, given input images $\boldsymbol{x} = \{x_i\}_{i=1}^{M}$ with their binary modality code $\boldsymbol{c} = \{c_i | c_i = 1 \text{ if } x_i \neq \mathbf{0} \text{ else } 0\}_{i=1}^{M}$, we first form a textual prompt following the template: "*The input MRI modalities are* {*available modalities*}." This prompt is fed into a frozen text encoder to obtain a textual semantic embedding $\boldsymbol{Z}_t$. Here, we employ the off-the-shelf CLIP text encoder (Radford et al. 2021) which has been pre-trained on a large amount of data and proven to be able to capture anatomical relationships (Liu et al. 2023b). Therefore, we use its prior knowledge to semantically model the inter-modal synergy for each specific missing modality combination by the extracted text embedding. Simultaneously, we apply global average pooling to the visual feature maps of our model, producing a feature-level representation $\boldsymbol{Z}_v$.

The textual and visual embeddings are concatenated to form a semantically-rich indicator, which is input into the hyper-network, i.e., a multi-layer perceptron (MLP), to generate the weights and biases for the set of $L_{hyp}$ prediction layers in the decoder as follows:

$$\{\boldsymbol{w}_i, \boldsymbol{b}_i\}_{i=1}^{L_{hyp}} = \texttt{MLP}(\boldsymbol{Z}_v \oplus \boldsymbol{Z}_t), \tag{6}$$

where $\boldsymbol{w}_i$ and $\boldsymbol{b}_i$ denote the weights and biases of the $i$-th adaptive layer. By fusing textual prompts and visual representations, our approach incorporates both explicit semantic knowledge of modality presence and contextual visual information. This enables our model to personalize for each specific input, effectively handling the heterogeneity in modality distribution caused by missing modalities.

## Overall Objective for Segmentation

For our segmentation network, the Dice loss is used to minimize the difference between the predicted segmentation mask and labels: $\mathcal{L}_{\text{task}} = \mathcal{L}_{\text{dice}}(\mathcal{S}(\texttt{MDrop}(x_a), \boldsymbol{c}), y)$, where $\boldsymbol{c}$ is the modality code indicating modality presence for prompt-guided modality adaptation. Overall, we jointly use segmentation loss, modality recovery model refining loss in Eq. (5), and KD loss in Eq. (4) to optimize our framework:

$$\min_{\theta_{\mathcal{S}}, \theta_{\mathcal{M}}} \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\mathcal{M}_{\text{rec}}} + \beta \mathcal{L}_{\text{KD}}, \tag{7}$$

where $\theta_{\mathcal{S}}$ represents the parameters of our student segmentation model, $\theta_{\mathcal{M}}$ is the parameters of the learnable modality recovering module, $\alpha$ and $\beta$ are used to control the refining of $\mathcal{M}_{\text{rec}}$ and knowledge distillation respectively.

# Experiment

## Datasets and Evaluation Metrics

We conduct all the experiments using BraTS2018 and BraTS2020 (Menze et al. 2015), two widely used datasets for multimodal brain tumor segmentation that comprise 285 and 369 preprocessed cases, respectively. Each subject has four MRI sequences in size of $155 \times 240 \times 240$. The segmentation labels include three classes, which are whole tumor (WT), tumor core (TC), and enhancing tumor (ET). We follow the data split in previous works (Liu et al. 2023a; Ding, Yu, and Yang 2021) for training, validation and testing, which is 199:29:57 for BraTS2018 and 219:50:100 for BraTS2020. Each MRI scan has been skull-stripped, co-registered to the same template, and re-sampled to 1 $\text{mm}^3$ resolution. Following the common practice for BraTS datasets, we adopt the Dice Similarity Coefficient (DSC ↑) and the 95th percentile Hausdorff Distance (HD95 ↓) to quantitatively evaluate the segmentation performance.

## Implementation Details

To simulate real-world heterogeneous training data, we construct the dataset by sampling a specified ratio of full-modal (FM) subjects and randomly discard 1–3 modalities in the rest. We adopt a 3D U-Net with group normalization as the backbone, following prior works (Wang et al. 2021; Liu et al. 2023a). Standard 3D augmentations including cropping, normalization, intensity transforms, and flipping are applied. Both pre-training and segmentation training use Adam optimizer (initial learning rate $3 \times 10^{-4}$, weight decay $5 \times 10^{-4}$), with cosine learning rate decay. Pre-training runs for 600 epochs (batch size 2), while segmentation employs early stopping (up to 1000 epochs, batch size 1). $\alpha$ and $\beta$ in Eq. (7) are set to 1.0 and 0.1 after a simple grid search. At inference, we evaluate our model under all 15 possible modality combinations and report the average.

## Comparison with State-of-the-art Methods

We compare our method with SOTA missing modality methods under the dual-stage missing modality scenario. RFNet (Ding, Yu, and Yang 2021), $\text{M}^3$AE (Liu et al. 2023a), and IM-Fuse (Pipoli et al. 2025) are designed to handle inference-stage missing modalities, whereas Meta-learning (Konwer et al. 2023) and $\text{M}^3$FeCon (Zeng et al. 2024) are for both training and inference stages. To enable the training of inference-stage methods, we use zero-padding to impute the absent training modalities. Since the code of Meta-learning is currently not available, we re-implement it on our backbone following the original paper and tune the hyperparameters. We train these methods using training sets with varying FM ratios (1%, 10%, and 50%), and then test them under all 15 different modality input scenarios. For a fair comparison, all these approaches are implemented using the same dataset split and backbone.

Tab. 1 shows the performance of our model and other SOTA methods on BraTS2018 under the dual-stage missing modality scenario. The FM ratio of training data is set to 1%, which challenges the model's robustness in handling training-stage missing modalities. Our method outperforms previous SOTA methods on the average DSC across all 15 inference-stage missing modality scenarios and three tumor regions. $\text{M}^3$AE is the second-best method overall, while our approach outperforms it by 4.84%, 1.86%, and 0.74% on ET, TC, and WT, respectively. The superior performance demonstrates that our model can effectively utilize the heterogeneous training data comprising both modality-complete and incomplete samples and robustly generate segmentation masks for inputs with different missing modality combinations during inference. We further set different FM ratios for the training data on BraTS2018 and BraTS2020,

| Region | | FLAIR | ○ | ○ | ○ | ● | ○ | ○ | ● | ○ | ● | ● | ● | ● | ● | ○ | ● | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T1 | ○ | ○ | ● | ○ | ○ | ● | ● | ● | ○ | ○ | ● | ● | ○ | ● | ● | |
| | | T1c | ○ | ● | ○ | ○ | ● | ● | ○ | ○ | ○ | ● | ● | ○ | ● | ● | ● | |
| | | T2 | ● | ○ | ○ | ○ | ● | ○ | ○ | ● | ● | ○ | ○ | ● | ● | ● | ● | |
| ET | RFNet | | 37.46 | 66.52 | 24.27 | 32.57 | 71.27 | 69.56 | 36.19 | 36.53 | 39.10 | 71.68 | 71.69 | 39.94 | 71.70 | 71.55 | 71.82 | 54.12* |
| | M$^3$AE | | 41.61 | 66.42 | 35.76 | 38.32 | 70.00 | 66.82 | 39.08 | 42.53 | 42.23 | 67.87 | 69.85 | 40.42 | 71.25 | 68.05 | 69.55 | 55.32* |
| | IM-Fuse | | 35.23 | 67.61 | 34.64 | 29.50 | 68.61 | 68.52 | 36.17 | 42.56 | 32.26 | 67.82 | 67.73 | 38.86 | 66.27 | 67.68 | 65.46 | 52.59* |
| | Meta | | 25.32 | 58.33 | 20.03 | 22.71 | 65.65 | 64.49 | 21.01 | 26.86 | 24.80 | 67.12 | 67.40 | 25.37 | 69.08 | 69.26 | 67.55 | 46.33* |
| | M$^3$FeCon | | 30.31 | 67.40 | 30.76 | 38.24 | 72.86 | 69.99 | 40.51 | 34.50 | 39.15 | 74.51 | 72.63 | 40.70 | 72.42 | 73.41 | 72.41 | 55.32* |
| | Ours | | **45.50** | **75.00** | **41.87** | **39.68** | **73.55** | **75.33** | **42.36** | **48.17** | **45.58** | **73.68** | **73.91** | **43.95** | **74.07** | **75.17** | **74.62** | **60.16** |
| TC | RFNet | | 63.55 | 77.42 | 54.81 | 57.60 | 79.74 | 79.59 | 65.42 | 66.83 | 66.86 | 79.04 | 79.71 | 69.38 | 79.97 | 79.85 | 80.16 | 72.00* |
| | M$^3$AE | | 68.48 | 80.72 | 67.64 | 67.14 | 82.21 | 80.96 | 70.45 | 70.26 | 70.81 | 82.18 | 82.33 | 70.94 | 82.82 | 82.29 | 82.49 | 76.11* |
| | IM-Fuse | | 63.18 | 76.95 | 61.85 | 62.81 | 78.83 | 78.99 | 69.89 | 68.10 | 67.75 | 79.40 | 80.93 | 70.40 | 80.51 | 79.45 | 80.70 | 73.32* |
| | Meta | | 50.00 | 70.87 | 47.04 | 47.83 | 75.19 | 71.96 | 52.64 | 54.57 | 52.33 | 74.84 | 73.98 | 57.49 | 74.99 | 76.41 | 74.79 | 63.66* |
| | M$^3$FeCon | | 53.42 | 71.47 | 56.03 | 61.67 | 78.09 | 78.58 | 68.68 | 59.49 | 65.39 | 81.66 | 81.99 | 68.63 | 81.32 | 80.67 | 81.56 | 71.24* |
| | Ours | | **73.29** | **83.33** | **68.19** | **69.10** | **84.56** | **83.50** | **71.23** | **73.39** | **73.02** | **83.67** | **83.62** | **73.16** | **83.48** | **83.11** | **82.89** | **77.97** |
| WT | RFNet | | 84.20 | 71.87 | 70.49 | 84.04 | 84.92 | 75.33 | 85.31 | 84.61 | 88.20 | 87.82 | 87.31 | 88.89 | **89.49** | 85.01 | **89.39** | 83.79* |
| | M$^3$AE | | 84.48 | **75.96** | 76.86 | **88.49** | 83.93 | **78.04** | **88.50** | 84.45 | 89.20 | 87.22 | 87.43 | 89.05 | 87.40 | 84.05 | 87.63 | 84.85* |
| | IM-Fuse | | 83.36 | 74.69 | 75.17 | 85.43 | 84.43 | 78.25 | 86.95 | 83.98 | 87.72 | 87.36 | 87.17 | 87.66 | 88.28 | 84.34 | 87.63 | 84.16* |
| | Meta | | 77.12 | 66.90 | 62.65 | 82.15 | 77.59 | 70.82 | 82.24 | 74.69 | 83.46 | 83.61 | 83.14 | 84.69 | 84.32 | 77.66 | 83.15 | 78.28* |
| | M$^3$FeCon | | 79.57 | 66.29 | 69.41 | 85.77 | 79.36 | 74.35 | 88.10 | 82.65 | 87.83 | 86.65 | 87.32 | 88.45 | 87.19 | 82.70 | 87.31 | 82.20* |
| | Ours | | **85.95** | 75.62 | **77.05** | 87.96 | **86.35** | 77.13 | 88.23 | **86.73** | **89.37** | **87.84** | **88.38** | **89.43** | 88.60 | **86.25** | 88.92 | **85.59** |

Table 1: Comparison with state-of-the-art methods on three tumor regions (ET, TC, WT) under 1% FM ratio on BraTS2018. DSCs across 15 inference-stage input modality scenarios are reported. The best and second best scores are **bolded** and underlined, respectively. Statistically significant improvements of average DSC with $p < 0.05$ are denoted by $*$.

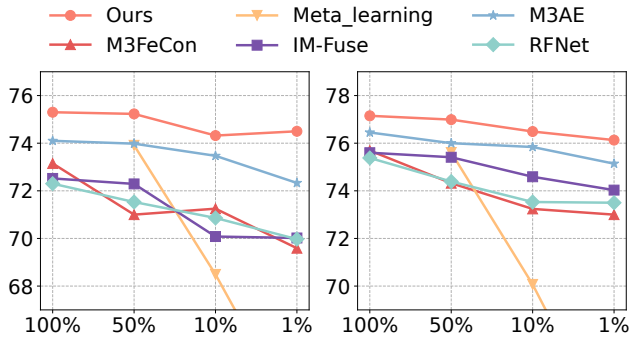


Figure 3: Comparison with SOTA methods on BraTS2018 (left) and BraTS2020 (right) under different FM ratios.

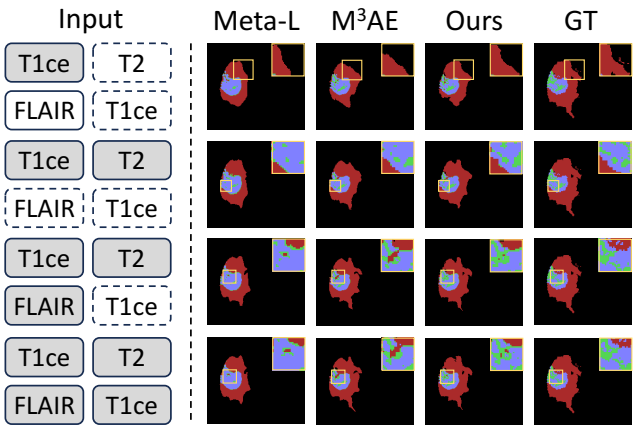


Figure 4: Visualization results from three methods (1% FM) for different input modality combinations. Red, green, and purple regions indicate WT, TC, and ET, respectively.

as shown in Fig. 3. Our method outperforms SOTA methods under all FM ratios, demonstrating its general robustness to varying degrees of missing modalities in training. Results on the HD95 metric are provided in the appendix.

Fig. 4 visualizes segmentation masks from Meta-learning, M$^3$AE, and our method with four different test-time input modality combinations. All the models are trained under 1% FM. As the number of input modalities decreases, our model consistently produces more accurate masks, showing superior effectiveness in dual-stage missing modality scenarios.

## Ablation Studies

**Effectiveness of key components** We first assess the contribution of each component in our framework: distribution approximation (DA), prompt-guided modality adaptation (PGMA), and complete-then-distill (CTD). Mask-reconstruciton pre-training serves as the baseline: we only calculate the reconstruction loss on available modalities of a specific sample; modality dropout is used to adapt it on the segmentation task. The dual-stage missing modality results (1% FM) in Tab. 2 show that incorporating DA during pre-training improves the average performance by 1.43% over the baseline. This demonstrates that DA can effectively compensate for the incomplete supervision caused by the lack

| Methods | 1% FM | | | | 100% FM | | | |
|---|---|---|---|---|---|---|---|---|
| | ET | TC | WT | Avg. | ET | TC | WT | Avg. |
| Mask-Rec | 55.53 | 75.53 | 84.97 | 72.01 | 60.07 | 77.22 | 85.30 | 74.20 |
| + DA | 58.34 | 76.99 | 84.99 | 73.44 | 60.69 | 78.05 | 85.43 | 74.72 |
| + PGMA | 59.76 | 77.68 | 85.51 | 74.32 | **61.31** | **78.91** | **85.69** | **75.30** |
| + CTD | **60.16** | **77.97** | **85.59** | **74.57** | — | — | — | — |

Table 2: Ablation study of key components under 1% and 100% FM ratio. DA, PGMA, and CTD refers to distribution approximation, prompt-guided modality adaptation, and complete-then-distill.

of reconstruction ground truth in missing modalities during pre-training, thereby enhancing the quality of modality-invariant representation learning. Based on this, we further incorporate the CTD to the segmentation model when fine-tuning the pre-trained model on segmentation tasks, aiming to enhance its capability of adapting to heterogeneous modality input combinations. This results in a 2.31% improvement compared to the baseline. By further introducing PGMA, we effectively utilize the recovered full-modal information to enhance partial-modal representations, and finally achieve a 2.56% advantage relative to baseline.

Tab. 2 further shows that DA and PGMA are effective for the inference-stage missing modality problem (100% FM). Here, CTD is not discussed as it is not required when training data are modality-complete. The results indicate that our DA improves both pre-training with incomplete data and general modality-invariant representation learning, as it guides the model to focus on important tumor region-related information by encouraging better segmentation performance from reconstructed modalities.

**Detailed analysis of the complete-then-distill paradigm** We conducted experiments using four configurations. Results are shown in Tab. 3. (a) is the baseline, which is the same as the Mask-Rec + DA in Tab. 2; (b) and (c) use the frozen teacher model to conduct feature-level knowledge distillation. Wherein, (b) merely uses teacher features of available modalities to enhance partial-modal representations in the student model. (c), on the other hand, adopts $\mathcal{M}_{\text{rec}}$ to complete missing modalities and extracts full-modal teacher features for distillation. (d) allows $\mathcal{M}_{\text{rec}}$ to be learnable during training to further refine modality completion quality. The comparison between (a) and (b) shows that using teacher features extracted from all available real modalities for knowledge distillation is ineffective and can slightly reduce performance, even though the teacher uses more modalities than the student's partial-modal features. This is due to the heterogeneity of modality distributions in the training data, which causes forceful alignment to lead to low-quality features and bias in the student model. Therefore, as shown in (c), introducing $\mathcal{M}_{\text{rec}}$ in complete-then-distill alleviates the information loss caused by missing modalities. The completed full-modal data enables the teacher to extract informative full-modal knowledge, which is then used to enhance the partial-modal representations in the student. (d) demonstrates the effectiveness of continuously refining $\mathcal{M}_{\text{rec}}$ during training with $\mathcal{L}_{\text{rec}}$ to improve the estimation quality of missing modalities.

| id | $\mathcal{M}_{\text{rec}}$ | KD | $\mathcal{L}_{\mathcal{M}_{\text{rec}}}$ | ET | TC | WT | Avg. |
|---|---|---|---|---|---|---|---|
| (a) | | | | 58.34 | 76.99 | 84.99 | 73.44 |
| (b) | | ✔ | | 58.35 | 76.79 | 85.09 | 73.41 |
| (c) | ✔ | ✔ | | 59.21 | 77.21 | **85.17** | 73.87 |
| (d) | ✔ | ✔ | ✔ | **59.61** | **78.10** | 85.05 | **74.25** |

Table 3: Analysis of prompt-guided modality adaptation. Results are reported by DSC under 1% FM ratio.

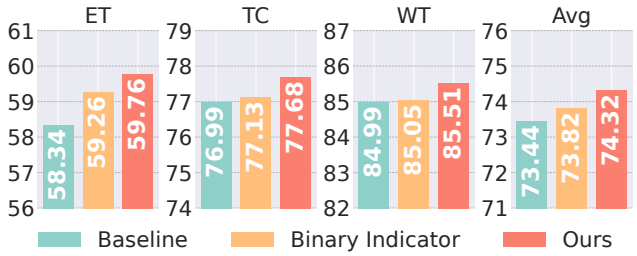


Figure 5: Analysis of prompt-guided modality adaptation.

**Detailed analysis of prompt-guided modality adaptation** Fig. 5 shows the ablation results of our prompt-guided modality adaptation. The baseline remains the same as (a) in Tab. 3. Experiments are conducted under 1% FM ratio, and average DSCs across 15 modality combinations are reported. We replace the text encoder with a 2-layer MLP to generate modality embedding using binary modality indicator. It can be seen that an improvement of 0.38% on the mean of three regions is achieved by using simple binary indicator, proving the effectiveness of adaptive model personalization using hyper-network. Nevertheless, binary codes assume that each modality is independent to other modalities and ignore the inter-modal correlations and synergies. By designing a more informative visual-textual indicator, our prompt-guided modality adaptation improves the performance on all three tumor regions, and the average DSC increases from 73.44% to 74.32%.

## Conclusion

In this paper, we present a universal model to tackle missing modalities during both training and inference. We innovate distribution approximation for mask-reconstruction pre-training, through which our model learns modality-invariant representations and missing modality reconstruction. After that, when adapting our model to segmentation tasks, we propose complete-then-distill paradigm to effectively utilize the full-modal knowledge recovered by reconstruction to enhance partial-modal representations in the student model. We also design a prompt-guided modality adaptation to alleviate modal distribution heterogeneity. Through robust modality reconstruction and model personalization, our model demonstrates superior performance on two brain tumor segmentation benchmarks, surpassing other methods under all missing stages with various missing ratios.

## Acknowledgments

This work was supported by Ministry of Education Tier 1 grant, Singapore (24-1250-P0001), and Ministry of Education Tier 2 grant, Singapore (T2EP20224-0028).

## References

Azad, R.; Khosravi, N.; and Merhof, D. 2022. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. In *Proc. Int. Conf. on Med. Imaging with Deep Learn.*, 48–62. PMLR.

Chen, C.; Dou, Q.; Jin, Y.; Chen, H.; Qin, J.; and Heng, P.-A. 2019a. Robust multimodal brain tumor segmentation via feature disentanglement and gated fusion. In *Proc. MICCAI*, 447–456. Springer.

Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019b. Data-free learning of student networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3514–3522.

Ding, Y.; Yu, X.; and Yang, Y. 2021. RFNet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 3975–3984.

Dou, Q.; Liu, Q.; Heng, P. A.; and Glocker, B. 2020. Unpaired multi-modal segmentation via knowledge distillation. *IEEE Trans. Med. Imaging*, 39(7): 2415–2425.

Ha, D.; Dai, A. M.; and Le, Q. V. 2017. HyperNetworks. In *Proc. Int. Conf. Learn. Represent.*

Hu, M.; Maillard, M.; Zhang, Y.; Ciceri, T.; La Barbera, G.; Bloch, I.; and Gori, P. 2020. Knowledge distillation from multi-modal to mono-modal segmentation networks. In *Proc. MICCAI*, 772–781. Springer.

Isensee, F.; Jaeger, P. F.; Kohl, S. A.; Petersen, J.; and Maier-Hein, K. H. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods*, 18(2): 203–211.

Konwer, A.; Hu, X.; Bae, J.; Xu, X.; Chen, C.; and Prasanna, P. 2023. Enhancing modality-agnostic representations via meta-learning for brain tumor segmentation. In *Proc. IEEE Int. Conf. Comput. Vis.*, 21415–21425.

Lee, D.; Moon, W.-J.; and Ye, J. C. 2020. Assessing the importance of magnetic resonance contrasts using collaborative generative adversarial networks. *Nat. Mach. Intell.*, 2(1): 34–42.

Liu, H.; Nie, D.; Shen, D.; Wang, J.; and Tang, Z. 2022. Multimodal brain tumor segmentation using contrastive learning based feature comparison with monomodal normal brain images. In *Proc. MICCAI*, 118–127. Springer.

Liu, H.; Wei, D.; Lu, D.; Sun, J.; Wang, L.; and Zheng, Y. 2023a. M3AE: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, 1657–1665.

Liu, J.; Zhang, Y.; Chen, J.-N.; Xiao, J.; Lu, Y.; A Landman, B.; Yuan, Y.; Yuille, A.; Tang, Y.; and Zhou, Z. 2023b. Clip-driven universal model for organ segmentation and tumor detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 21152–21164.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; Lanczi, L.; Gerstner, E.; Weber, M.-A.; Arbel, T.; Avants, B. B.; Ayache, N.; Buendia, P.; Collins, D. L.; Cordier, N.; Corso, J. J.; Criminisi, A.; Das, T.; Delingette, H.; Demiralp, c.; Durst, C. R.; Dojat, M.; Doyle, S.; Festa, J.; Forbes, F.; Geremia, E.; Glocker, B.; Golland, P.; Guo, X.; Hamamci, A.; Iftekharuddin, K. M.; Jena, R.; John, N. M.; Konukoglu, E.; Lashkari, D.; Mariz, J. A.; Meier, R.; Pereira, S.; Precup, D.; Price, S. J.; Raviv, T. R.; Reza, S. M. S.; Ryan, M.; Sarikaya, D.; Schwartz, L.; Shin, H.-C.; Shotton, J.; Silva, C. A.; Sousa, N.; Subbanna, N. K.; Szekely, G.; Taylor, T. J.; Thomas, O. M.; Tustison, N. J.; Unal, G.; Vasseur, F.; Wintermark, M.; Ye, D. H.; Zhao, L.; Zhao, B.; Zikic, D.; Prastawa, M.; Reyes, M.; and Van Leemput, K. 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging*, 34(10): 1993–2024.

Mordvintsev, A.; Olah, C.; and Tyka, M. 2015. Inceptionism: Going deeper into neural networks. *Google research blog*, 20(14): 5.

Pereira, S.; Pinto, A.; Alves, V.; and Silva, C. A. 2016. Brain tumor segmentation using convolutional neural networks in MRI images. *IEEE Trans. Med. Imaging*, 35(5): 1240–1251.

Pipoli, V.; Saporita, A.; Marchesini, K.; Grana, C.; Ficarra, E.; and Bolelli, F. 2025. IM-Fuse: A Mamba-based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities. In *28th International Conference on Medical Image Computing and Computer Assisted Intervention*, 1–11. Springer.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proc. Int. Conf. Mach. Learn.*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ranjbarzadeh, R.; Bagherian Kasgari, A.; Jafarzadeh Ghoushchi, S.; Anari, S.; Naseri, M.; and Bendechache, M. 2021. Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images. *Sci. Rep.*, 11(1): 1–17.

Sharma, A.; and Hamarneh, G. 2019. Missing MRI pulse sequence synthesis using multi-modal generative adversarial network. *IEEE Trans. Med. Imaging*, 39(4): 1170–1183.

Wang, Y.; Zhang, Y.; Liu, Y.; Lin, Z.; Tian, J.; Zhong, C.; Shi, Z.; Fan, J.; and He, Z. 2021. Acn: Adversarial co-training network for brain tumor segmentation with missing modalities. In *Proc. MICCAI*, 410–420.

Wei, S.; Luo, C.; and Luo, Y. 2023. MMANet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 20039–20049.

Xing, Z.; Yu, L.; Wan, L.; Han, T.; and Zhu, L. 2022. NestedFormer: Nested modality-aware transformer for brain tumor segmentation. In *Proc. MICCAI*, 140–150. Springer.

Yang, H.; Sun, J.; and Xu, Z. 2023. Learning unified hypernetwork for multi-modal MR image synthesis and tumor segmentation with missing modalities. *IEEE Trans. Med. Imaging*.

Yun, S.; Choi, I.; Peng, J.; Wu, Y.; Bao, J.; Zhang, Q.; Xin, J.; Long, Q.; and Chen, T. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37: 98782–98805.

Zeng, Z.; Peng, Z.; Yang, X.; and Shen, W. 2024. Missing as Masking: Arbitrary Cross-Modal Feature Reconstruction for Incomplete Multimodal Brain Tumor Segmentation. In *Proc. MICCAI*, 424–433. Springer.

Zhang, D.; Huang, G.; Zhang, Q.; Han, J.; Han, J.; Wang, Y.; and Yu, Y. 2020. Exploring task structure for brain tumor segmentation from multi-modality MR images. *IEEE Trans. Med. Imaging*, 29: 9032–9043.

Zhang, Y.; He, N.; Yang, J.; Li, Y.; Wei, D.; Huang, Y.; Zhang, Y.; He, Z.; and Zheng, Y. 2022. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation. In *Proc. MICCAI*, 107–117. Springer.

Zhou, T.; Canu, S.; Vera, P.; and Ruan, S. 2021. Latent correlation representation learning for brain tumor segmentation with missing MRI modalities. *IEEE Trans. Image Process.*, 30: 4263–4274.