THEME: TASK-HIERARCHICAL EVIDENCE COMPLETION FOR INCOMPLETE MULTIMODAL BRAIN TUMOR SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

An unavailable MRI sequence removes information needed for brain tumor segmentation, but its effect differs across tumor regions. We propose THEME, which predicts a separate learned summary for each missing sequence and each of three region tasks. Modality-specific encoders extract spatial features and task summaries; a completion module estimates the summaries of missing sequences from those observed. One router fuses only acquired spatial features, and a second combines acquired and predicted summaries for each task before decoder modulation. We evaluate every one of the 15 non-empty subsets of four MRI sequences. On BraTS 2018, THEME obtains 88.44/81.16/64.32% Dice for whole tumor/tumor core/enhancing tumor, improving on the strongest baseline by 1.50/0.35/1.87 points. The corresponding BraTS 2020 scores are 89.65/83.04/68.16%, with gains of 1.49/0.36/1.87 points. Removing direct summary supervision reduces enhancing-tumor Dice by 5.78 points. These experiments show the benefit of the complete architecture under simulated missing-sequence evaluation.

1 INTRODUCTION

Brain tumor segmentation combines MRI sequences that reveal different aspects of the same lesion. FLAIR, T1, contrast-enhanced T1 (T1ce), and T2 support the delineation of enhancing tumor (ET), tumor core (TC), and whole tumor (WT), three nested regions used in BraTS [15; 4]. When one or more sequences are unavailable, the remaining images provide uneven support for these regions. A model may identify the broad tumor extent while struggling to separate its inner regions. Robust segmentation therefore requires attention to how each available sequence informs each target region. The challenge involves both localization and discrimination: identifying the extent of a lesion and distinguishing its internal regions use different combinations of image cues. Losing a sequence changes the information available for these decisions even though the underlying anatomy is unchanged. The input configuration is consequently part of the segmentation problem. A useful model must support examinations with different acquired sequences and adapt its regional predictions to the information present in each examination.

Existing approaches to incomplete MRI inputs commonly aggregate observed features, synthesize missing images, or complete a shared representation. These strategies address missing information, but the information needed for ET, TC, and WT differs. A single sequence can be useful for one region and less informative for another, so completion at a shared level leaves the region-specific role of that sequence implicit. We instead ask what feature summary a missing sequence would provide for each segmentation task, and how those summaries can be learned from the sequences that remain. This question becomes especially relevant when only one or two sequences are available. Two examinations with the same number of sequences can offer different support for the same region, and the same examination can support WT more strongly than ET. The nested target hierarchy provides a way to organize this uneven information: completion can be indexed by the region decision that will consume it. This motivates learning what is missing for each task while retaining the spatial information in the acquired images.

Our contributions are threefold. **First**, we formulate missing-sequence completion around a learned summary for each modality and region task, making the segmentation target part of the comple-

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.

tion problem. **Second**, we develop THEME to predict these summaries from available sequences and use them for region prediction alongside spatial features extracted from acquired images; we call the learned summaries "evidence." **Third**, we evaluate the complete system on all non-empty combinations of four MRI sequences in BraTS 2018 and 2020, and analyze its components and performance under sparse inputs. The evaluation combines repeated training runs, cumulative architecture ablations, individual loss removals, and a prostate segmentation study to examine how the design behaves across input configurations and target hierarchies.

2 RELATED WORK

Fusion with incomplete inputs. HeMIS combines modality-specific feature statistics to accommodate variable input sets [7]. Transformer approaches model interactions among acquired representations: mmFormer combines modality-aware features, M^2FTrans uses modality-masked fusion, and TMFormer merges tokens [24; 16; 25]. RFNet explicitly makes fusion depend on the tumor region, showing the relevance of regional conditioning for incomplete MRI segmentation [6]. ShaSpec separates shared and specific features using distribution alignment and domain classification, then combines them through residual fusion [18]. MoMKE further organizes multimodal learning around modality experts [21]. These approaches provide mechanisms for selecting and combining information available in the input. THEME uses modality experts for spatial maps and adds a separately supervised completion target for every missing modality–region summary.

Reconstruction and regional representations. Generative approaches recover missing information at the image or feature level. Hyper-GAE learns a unified hyper-network for synthesis and tumor segmentation, while diffusion-based methods model missing sequences through conditional generation [22; 1]. RA-HVED incorporates region-of-interest attention into heteromodal variational completion [8]. RMBTS uses feature disentanglement and gated fusion to learn representations for incomplete brain tumor inputs [5]. These designs connect reconstruction or representation learning with the segmentation target. The object predicted by THEME is a compact feature summary indexed by both modality and region. Its supervision is supplied by the corresponding full-sequence expert, and the completed summary enters the task branch of the decoder.

Learning across modality subsets. Full-input supervision can guide models operating on partial inputs. ACN couples full and missing-modality networks through adversarial co-training [20]. LCKD learns cross-modal knowledge transfer, and AdaMM adapts segmentation through knowledge distillation [19; 29]. M3AE combines masked multimodal representation learning with self-distillation [14]. SMML and UniMRSeg exploit relationships among partial views and hierarchical targets, whereas SimMLM compares richer and poorer modality inputs [13; 26; 11]. Konwer et al. use meta-learning to enrich partial representations with limited full-modality training data [9]. SIUM models uncertainty across input sets [3]. These training strategies act on predictions or learned representations. THEME combines segmentation supervision across two nested input sets with direct reconstruction of hidden modality–task summaries.

Task conditioning and feature completion. The completion unit determines how a model can use predicted information. Image synthesis produces voxel intensities that are encoded again. A shared latent predictor estimates a feature used by the segmentation model as a whole. In THEME, the missing unit is a modality–task summary $z_{u,t}$ extracted by the corresponding full-sequence expert. This unit has an explicit target during training and a task-specific consumer during inference. Spatial detail is supplied by the maps of the acquired sequences; completion supplies compact context for decoder modulation. This connects the completion objective directly to the representation consumed by the task branch.

Relationship to token fusion and cross-view training. TMFormer merges spatially redundant tokens within modalities and fuses tokens from present modalities. Its token merging operates on the available input representation. THEME additionally estimates the summaries of unavailable sequences and supervises them against full-sequence summaries. SMML aligns masked views with pixel-level and feature-level consistency, while SimMLM trains richer and poorer modality views. Our richer–poorer objective complements a direct modality–task reconstruction loss. The distinction is the object reconstructed and where that object enters the decoder: each missing sequence has

three supervised summaries, and those summaries influence task fusion separately from observed spatial fusion.

3 METHOD

THEME processes an incomplete MRI examination in four steps (Figure 1). Each available sequence first produces a spatial feature map and three region-specific feature summaries. A completion module then predicts the summaries of missing sequences. Two routing modules combine the spatial maps and the summaries, respectively. Finally, each fused summary controls a spatial gate in the segmentation head. We use "evidence" to denote these learned modality–region summaries throughout the paper.

3.1 INPUT SEQUENCES AND REGION TASKS

Let $\mathcal{M} = \{\text{FLAIR}, \text{T1}, \text{T1ce}, \text{T2}\}$ be the four MRI modalities and $\varnothing \neq \mathcal{A} \subseteq \mathcal{M}$ the modalities available for one subject. The input is $\{x_m : m \in \mathcal{A}\}$; the output consists of voxel-wise probabilities for ET, TC, and WT. To index the summary and gating branches, we decompose the nested labels into three exclusive regions:

$$R_1 = \text{ET}, \qquad R_2 = \text{TC} \setminus \text{ET}, \qquad R_3 = \text{WT} \setminus \text{TC}. \tag{1}$$

Task $t \in \{1, 2, 3\}$ refers to the branch associated with R_t. Thus R_t specifies an anatomical label region, $z_{m,t} \in \mathbb{R}^{64}$ will summarize modality m for that region, and the final predictions will be supervised on the three nested benchmark labels.

3.2 EXTRACTING MODALITY–REGION SUMMARIES

Each modality has an independent 3D encoder–decoder expert with the same topology. An observed sequence x_m produces a decoded spatial map $F_m \in \mathbb{R}^{C \times H \times W \times D}$ and a deepest encoder map H_m. Spatial fusion uses F_m, whereas summary extraction pools H_m. For encoder-grid location i, let $h_{m,i}$ denote its feature vector and $M_{m,t,i}$ the region mask resized to that grid. During training, this mask comes from the region labels in Eq. 1. During inference, the expert's segmentation head supplies

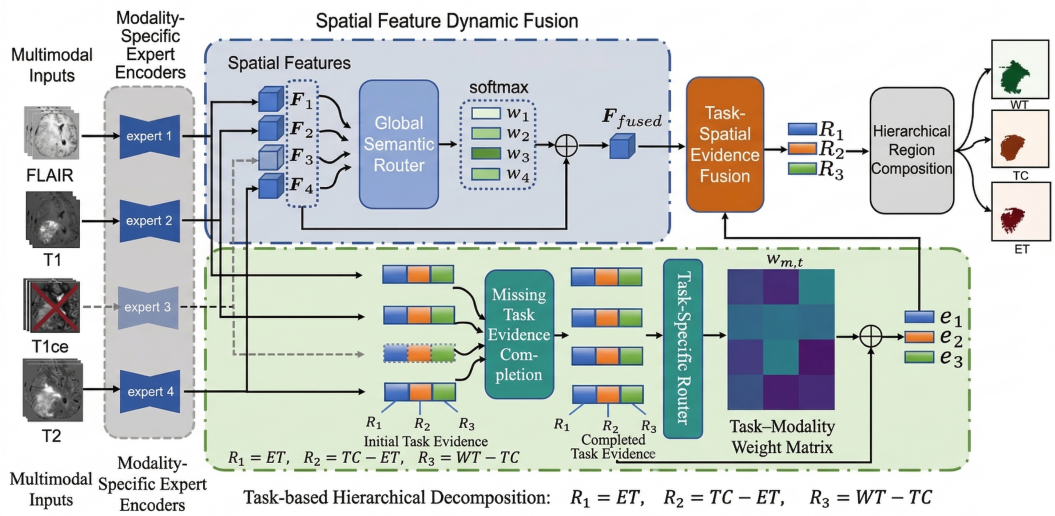

Figure 1: Data flow in THEME. Each available sequence supplies a spatial feature map and three region-specific summaries. Missing summaries are predicted from observed summaries. The global router combines acquired maps; the task router combines summaries. Each task summary controls a spatial gate before the three output logits are combined into ET, TC, and WT predictions.

detached probabilities: ET for R_1, the clipped TC–ET difference for R_2, and the clipped WT–TC difference for R_3.

For each region, a learned query $p^{\text{ext}}_{m,t} \in \mathbb{R}^d$ with $d = 64$ scores the encoder features. Two learned score channels, indexed by $k \in \{1, 2\}$, extract two pooled tokens:

$$\begin{aligned} a_{m,t,k,i} &= \mathrm{softmax}_i\left(s_{m,t,k}(h_{m,i}) + \frac{(p^{\text{ext}}_{m,t})^\top W^K_{m,t} h_{m,i}}{\sqrt{d}} + \log\max(M_{m,t,i}, \epsilon) \right), \\ v_{m,t,k} &= W^V_{m,t} \sum_i a_{m,t,k,i} h_{m,i}, \qquad z_{m,t} = f_{m,t}([v_{m,t,1}; v_{m,t,2}]). \end{aligned} \tag{2}$$

Here $s_{m,t,k}$ is a $1 \times 1 \times 1$ convolutional score channel, $W^K_{m,t}$ and $W^V_{m,t}$ are learned projections, $f_{m,t}$ is an MLP, $\epsilon = 10^{-6}$, and $[\cdot;\cdot]$ denotes concatenation. The softmax normalizes over spatial locations. The query supplies a region-specific feature score, while the mask biases pooling toward that region. Aggregating the two tokens produces one 64-dimensional summary $z_{m,t}$. Each observed modality therefore contributes one spatial map and three summary vectors.

3.3 Completing summaries of missing sequences

For every missing modality $u \in \mathcal{M} \setminus \mathcal{A}$, the completion module predicts $\hat{z}_{u,t}$ from $\{z_{m,t} : m \in \mathcal{A}\}$. A modality embedding $q_m \in \mathbb{R}^d$ identifies each sequence, and a learned completion query $p^{\text{cmp}}_t \in \mathbb{R}^d$ identifies each region. These parameters are separate from the extraction queries. Define $r_{u,t} = Q_t([q_u; p^{\text{cmp}}_t])$, $k_{m,t} = K_t([z_{m,t}; q_m])$, and $v_{m,t} = V_t([z_{m,t}; q_m])$, where Q_t, K_t, V_t are learned MLPs. Cross-modal attention selects the observed summaries relevant to target modality u:

$$c_{u,t} = P_t\left(\sum_{m \in \mathcal{A}} \frac{\exp(r^\top_{u,t} k_{m,t}/\sqrt{d})}{\sum_{j \in \mathcal{A}} \exp(r^\top_{u,t} k_{j,t}/\sqrt{d})} v_{m,t} \right), \qquad b_t = \frac{1}{|\mathcal{A}|} \sum_{m \in \mathcal{A}} z_{m,t}. \tag{3}$$

P_t projects the attention context, and b_t is the mean observed summary for region t. The predictor starts from b_t and adds a modality-specific residual. Its three MLP heads are evaluated in the order $R_1 \to R_2 \to R_3$:

$$\begin{aligned} \delta_{u,t} &= h_t([c_{u,t}; b_t; q_u; \rho_{u,<t}]), \qquad \rho_{u,t} = b_t + \delta_{u,t}, \\ \hat{z}_{u,t} &= b_t + \eta_t \delta_{u,t}, \qquad t = 1, 2, 3. \end{aligned} \tag{4}$$

The first head has no preceding ρ input; the second receives $\rho_{u,1}$, and the third receives both $\rho_{u,1}$ and $\rho_{u,2}$. η_t is a learned residual scale initialized to one. This ordered conditioning lets the outer-region predictor use the estimates for the inner regions. The tensor passed to task fusion retains each observed summary and fills each missing slot with a weighted completion:

$$\tilde{z}_{m,t} = \begin{cases} z_{m,t}, & m \in \mathcal{A}, \\ 0.7 b_t + 0.3 \hat{z}_{m,t}, & m \in \mathcal{M} \setminus \mathcal{A}. \end{cases} \tag{5}$$

3.4 Two routing modules and spatial modulation

The global router supplies spatial information shared by all three tasks. It scores the acquired maps using their pooled features, $g_m = \mathrm{GAP}(F_m)$, and computes

$$\alpha_m = \frac{\exp(s_m)}{\sum_{j \in \mathcal{A}} \exp(s_j)}, \qquad F = \sum_{m \in \mathcal{A}} \alpha_m F_m. \tag{6}$$

The score s_m is learned from g_m, modality identity, and availability, with a learned uncertainty penalty; Appendix A gives its parameterization. This branch obtains its spatial maps from the available sequences.

The task router supplies a different modality mixture for each region. For learned task scores $w_{m,t}$, it computes

$$\alpha_{m,t} = \frac{\exp(w_{m,t})}{\sum_{j \in \mathcal{M}} \exp(w_{j,t})}, \qquad \bar{z}_t = \sum_{m \in \mathcal{M}} \alpha_{m,t} \tilde{z}_{m,t}. \tag{7}$$

Each score uses $\tilde{z}_{m,t}$, a separate routing query p_t^{route}, modality identity, and the global score. Thus the two routing modules have explicit roles: α_m combines spatial maps into F, and $\alpha_{m,t}$ combines summaries into $\bar{z}_t$ for region t.

To apply $\bar{z}_t$ spatially, a linear projection is broadcast over the grid and concatenated with F. A convolutional head followed by sigmoid produces a one-channel gate G_t; a separate convolutional transform T_t produces a feature residual:

$$G_t = \sigma(\phi_t([F; \text{broadcast}(U_t \bar{z}_t)])), \quad F_t = F + \gamma_t G_t \odot T_t(F), \quad O_t = C_t(F_t). \tag{8}$$

γ_t is a learned scalar, $\odot$ denotes element-wise multiplication with channel broadcasting, and C_t is a one-channel prediction head. The summary determines where the branch applies its feature residual. The heads supply additive logits for the nested outputs:

$$\ell_{\text{ET}} = O_1, \qquad \ell_{\text{TC}} = O_1 + O_2, \qquad \ell_{\text{WT}} = O_1 + O_2 + O_3. \tag{9}$$

The final region probabilities are $\sigma(\ell_{\text{ET}})$, $\sigma(\ell_{\text{TC}})$, and $\sigma(\ell_{\text{WT}})$. The exclusive regions in Eq. 1 define pooling masks and gate supervision; Eq. 9 defines the segmentation outputs through cumulative logit addition.

3.5 Training objectives and inference

Training subjects contain all four sequences. We sample two nested, non-empty availability sets $S^- \subset S^+ \subseteq \mathcal{M}$ for the same subject and form two partial inputs. A full-sequence pass extracts reference summaries $z_m^\star = \text{sg}([z_{m,1}; z_{m,2}; z_{m,3}])$, where sg stops gradients. For view S, let $\hat{z}_m(S) = [\hat{z}_{m,1}; \hat{z}_{m,2}; \hat{z}_{m,3}]$. Reconstruction averages over its hidden modalities:

$$\mathcal{L}_{\text{evi}}(S) = \frac{1}{|\mathcal{M} \setminus S|} \sum_{m \in \mathcal{M} \setminus S} \left[\frac{\|\hat{z}_m(S) - z_m^\star\|_1}{3d} + 0.1\big(1 - \cos(\hat{z}_m(S), z_m^\star)\big) \right]. \tag{10}$$

It is zero for a full-input view. The reconstruction term supervises the predicted summaries before the weighted fusion in Eq. 5.

Segmentation minimizes the mean Dice and binary cross-entropy (BCE) losses over ET, TC, and WT in both partial views. Write $D_r(S)$ and $B_r(S)$ for these losses on output region r. Output ranking compares the two views for each region and loss term; evidence ranking compares their completed summary tensors with the full-input reference:

$$\begin{aligned} \mathcal{L}_{\text{rank}}^{\text{out}} &= \frac{1}{3} \sum_{r \in \{\text{ET,TC,WT}\}} \big([D_r(S^+) - D_r(S^-)]_+ + [B_r(S^+) - B_r(S^-)]_+ \big), \\ D_z(S) &= \frac{1}{4 \cdot 3d} \sum_{m \in \mathcal{M}} \|\tilde{z}_m(S) - z_m^\star\|_1, \qquad \mathcal{L}_{\text{rank}}^{\text{evi}} = [D_z(S^+) - D_z(S^-)]_+. \end{aligned} \tag{11}$$

Here $[a]_+ = \max(0, a)$ and $\tilde{z}_m$ concatenates the three completed summaries. Both partial views receive gradients from the ranking terms; the full-input reference is detached.

Gate supervision uses the resized region masks Y_t from Eq. 1, with task weights $w = (2, 1, 1)$. The regularizers are

$$\begin{aligned} \mathcal{L}_g &= \frac{1}{2} \sum_t \frac{w_t}{\sum_s w_s} \left[\text{BCE}(G_t, Y_t) + 1 - \text{Dice}_{\text{soft}}(G_t, Y_t) \right], \\ \mathcal{L}_a &= \frac{1}{3} \sum_t \text{mean}(G_t), \qquad \mathcal{L}_d = \sum_{s<t} \left[0.15 - \frac{1}{4} \sum_{m \in \mathcal{M}} |\alpha_{m,s} - \alpha_{m,t}| \right]_+ . \end{aligned} \tag{12}$$

These terms align each gate with its region, penalize gate area, and separate task-wise modality weights. We define $\mathcal{L}_{\text{reg}} = \lambda_g \mathcal{L}_g + \lambda_a \mathcal{L}_a + \lambda_d \mathcal{L}_d$ and optimize

$$\mathcal{L} = \mathcal{L}_{\text{seg}} + \lambda_o \mathcal{L}_{\text{rank}}^{\text{out}} + \lambda_e \mathcal{L}_{\text{evi}} + \lambda_r \mathcal{L}_{\text{rank}}^{\text{evi}} + \mathcal{L}_{\text{reg}}. \tag{13}$$

Losses are averaged over subjects, and reconstruction is averaged over the missing slots of the paired batch. Appendix A.1 gives the weights and gate-mask processing. At inference, the model runs the experts for $m \in \mathcal{A}$, completes the missing summary slots, and performs routing and prediction once. The label masks and full-input reference pass are used only during training.

Method	Year	BraTS 2018: all 15 sets WT	TC	ET	BraTS 2020: all 15 sets WT	TC	ET	$\lvert\mathcal{A}\rvert \leq 2$: TC/ET avg. B18	B20
LCKD [19]	2023	77.23 ±0.38	69.62 ±0.63	49.60 ±1.02	78.48 ±0.39	71.51 ±0.64	53.44 ±1.02	55.83 ±0.91	58.72 ±0.89
mmFormer [24]	2022	84.19 ±0.27	75.06 ±0.52	55.15 ±0.94	85.49 ±0.34	76.96 ±0.62	58.99 ±0.97	61.44 ±0.78	64.08 ±0.83
M^{2}FTrans [16]	2024	85.04 ±0.29	77.64 ±0.47	58.13 ±0.88	86.34 ±0.29	79.52 ±0.54	61.98 ±0.89	64.08 ±0.69	67.21 ±0.71
MSTKDNet [28]	2025	79.86 ±0.42	75.13 ±0.68	56.57 ±1.07	81.11 ±0.43	77.02 ±0.69	60.41 ±1.07	62.21 ±0.89	65.11 ±0.90
AdaMM [29]	2026	86.94 ±0.23	80.81 ±0.44	62.45 ±0.79	88.16 ±0.24	82.68 ±0.43	66.29 ±0.78	68.32 ±0.58	71.36 ±0.59
MoMKE [21]	2024	86.22 ±0.31	78.58 ±0.56	55.09 ±0.93	87.47 ±0.33	80.53 ±0.59	58.93 ±0.93	64.09 ±0.73	66.74 ±0.76
SimMLM [11]	2025	86.28 ±0.26	80.24 ±0.49	56.70 ±0.85	87.53 ±0.27	82.14 ±0.48	60.54 ±0.84	65.75 ±0.64	68.41 ±0.66
THEME	—	**88.44 ±0.21**	**81.16 ±0.39**	**64.32 ±0.73**	**89.65 ±0.21**	**83.04 ±0.38**	**68.16 ±0.73**	**70.22 ±0.53**	**73.27 ±0.55**

Table 1: Dice (%) across five runs. B18 and B20 denote BraTS 2018 and 2020. The WT, TC, and ET columns average over all 15 non-empty input sets in each dataset. The last two columns average TC and ET over the ten sets with at most two available sequences. Each cell gives the mean (upper line) and ± standard deviation (lower line) of five per-run subset averages.

4 EXPERIMENTS

4.1 PROTOCOL

BraTS 2018 and 2020 supply registered volumes for the four sequences and voxel annotations for WT, TC, and ET. For each trained model we test every non-empty $\mathcal{A} \subseteq \mathcal{M}$. The ten subsets with $|\mathcal{A}| \leq 2$ form the severe-missing group. We average Dice over the stated subset pool for each of five independently trained models, then report mean ± standard deviation across those five averages. A supplementary matrix instead reports variation across input sets for each method profile; the two types of spread have different denominators. All main-table methods use the same split and 15-subset evaluation. Appendix A.2 gives preprocessing, training settings, and the full subset definition.

Preprocessing and optimization. Each MRI sequence is normalized independently to zero mean and unit variance over non-zero foreground voxels; background remains zero. Training uses $128 \times 128 \times 128$ patches, padding before random cropping. Validation uses center cropping followed by padding. The four sequences of a subject receive the same spatial transform, preserving their voxel correspondence after cropping and flipping. Flips are applied independently along each axis with probability 0.5. Gaussian noise, Gaussian smoothing, contrast adjustment, and intensity scaling each have probability 0.15. These operations are shared by the variants in the ablations.

Optimization uses AdamW with weight decay 10^{-2}, an initial learning rate of 10^{-3}, and a cosine schedule reaching 10^{-6} over 300 epochs. Batch size counts two subject pairs, each providing a richer and a poorer input in the same update. Each pair uses two nested, non-empty availability sets from the same subject, so both views share the anatomy and segmentation labels while differing in the sequences supplied to the model. Dice and BCE supervise both views. A full-sequence pass supplies detached targets for their missing modality–task summaries; its segmentation output is not used at inference. Direct completion supervision ramps during the first 20 epochs, and the optional prototype anchor is disabled. The encoder, completion predictor, routers, and decoder are jointly trained. At inference only available sequences are encoded, and each missing summary is computed once before task routing.

Training and evaluation units. An availability configuration specifies which of the four sequences is supplied; it does not train a separate model. The same model is evaluated with each of the 15 input sets. Four sets contain one sequence, six contain two, four contain three, and one contains all four. Removing channels at evaluation leaves the remaining registered images and ground-truth segmentation unchanged, so differences across input sets measure the response to the information

supplied for the same subjects. The two BraTS releases are evaluated separately with the same target definition and training procedure.

For five-seed reporting, let $D_{k,t}(S)$ be the Dice of seed k for target t and availability set S. We first form $\bar{D}_{k,t} = |\mathcal{S}|^{-1} \sum_{S \in \mathcal{S}} D_{k,t}(S)$, then summarize the five $\bar{D}_{k,t}$ values. For the severe-missing columns, $\mathcal{S}$ contains the ten sets with one or two sequences, and each seed averages TC and ET before the five-run summary. Thus each availability set has equal weight, and the five-run standard deviation measures training variation after subset averaging. The complete tables retain the subset cells to expose different responses to, for example, T1ce-only and FLAIR-only input.

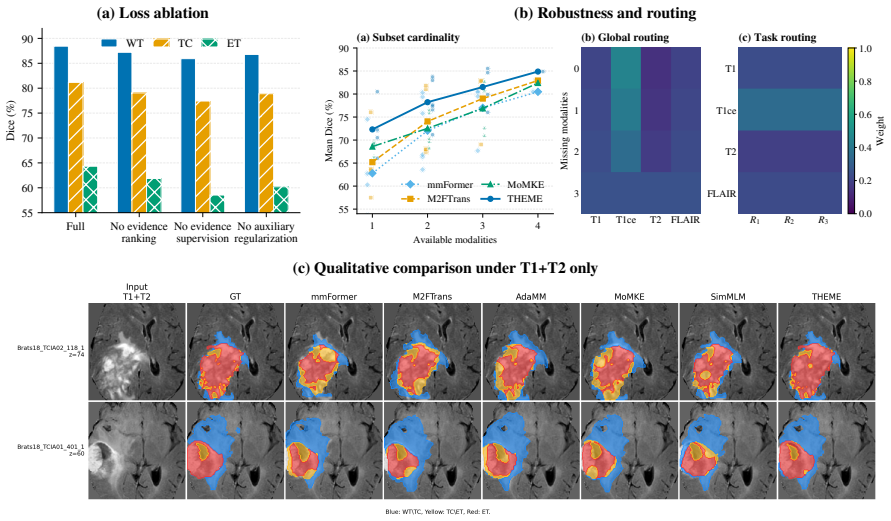

Figure 2: **Ablation, robustness, and qualitative diagnostics on BraTS.** (a) WT, TC, and ET loss-component ablations averaged over the 15 BraTS 2018 subsets; (b) Subset-cardinality performance and mean global/task routing weights. (c) Two severe-missingness cases with T1+T2 only; blue, yellow, and red denote WT\TC, TC\ET, and ET.

4.2 Segmentation Results

Table 1 shows that THEME improves over AdaMM by 1.50/0.35/1.87 Dice points for WT/TC/ET on BraTS 2018 and 1.49/0.36/1.87 points on BraTS 2020. The TC difference is small relative to the reported run-to-run standard deviations, so our primary accuracy findings are the WT and ET gains. For the ten severe-missing input sets, the mean TC/ET Dice is 70.22% on BraTS 2018 and 73.27% on BraTS 2020, exceeding AdaMM by 1.90 and 1.91 points, respectively. The grouped columns report all-set Dice and restricted-set TC/ET averages for both releases. Per-set results in the appendix show that the identity of the available sequences matters in addition to their count.

4.3 Component and Loss Analysis

Table 2 builds the architecture in four steps. Each row retains the preceding modules. Task evidence, missing-evidence completion, dual routing, and the complete training objective successively improve WT, TC, and ET on both BraTS releases.

Figure 2 shows that Dice generally rises with the number of acquired sequences, although sequence identity also matters. Global and task routing

Variant	BraTS 2018			BraTS 2020		
	WT	TC	ET	WT	TC	ET
Baseline	67.64	61.64	59.41	68.71	63.12	62.75
Baseline + TE	76.57	68.93	60.77	77.83	70.56	64.41
Baseline + TE + EC	80.95	74.04	62.18	82.18	75.96	65.83
Baseline + TE + EC + DR	84.87	77.61	63.40	86.06	79.72	67.01
THEME	**88.44**	**81.16**	**64.32**	**89.65**	**83.04**	**68.16**

Table 2: Cumulative BraTS 2018/2020 ablation; Dice (%) over 15 input sets. TE: task evidence; EC: completion; DR: dual routing.

have different modality-weight profiles, and the three tasks assign different weights. The global route mixes observed spatial features; the task route selects among observed and completed summaries.

Removing direct completion supervision has the largest effect in Table 3: ET falls by 5.78 points on BraTS 2018 and 6.07 points on BraTS 2020. Removing evidence ranking reduces ET by 2.45 and 2.73 points, respectively; removing auxiliary regularization reduces ET by 4.11 and 4.22 points. Each row disables one loss group from the complete model.

Variant	BraTS 2018			BraTS 2020		
	WT	TC	ET	WT	TC	ET
Full THEME	**88.44**	**81.16**	**64.32**	**89.65**	**83.04**	**68.16**
No direct supervision	85.93	77.42	58.54	87.16	79.37	62.09
No evidence ranking	87.21	79.05	61.87	88.36	81.02	65.43
No auxiliary regularization	86.78	78.89	60.21	88.02	80.63	63.94

Table 3: Loss ablation on BraTS 2018 and 2020; Dice (%).

The qualitative figure in Appendix B shows T1+T2 input because both contrast-enhanced T1 and FLAIR are absent, making inner-region errors easy to inspect. It is a case illustration on BraTS 2018; Table 1 provides the dataset-level comparison for both releases.

4.4 CROSS-TASK SEGMENTATION

Prostate158 changes the organ, target hierarchy, and sequence set: a cancer lesion is nested within the prostate gland, while the inputs are T2W/DWI/ADC. Table 4 aggregates Dice over all seven non-empty availability subsets. THEME reaches 46.80% lesion Dice and 88.90% gland Dice (67.90% average), compared with 44.00/88.20/66.10% for its no-evidence counterpart and 43.10/87.80/65.50% for MoMKE, the strongest listed baseline. This cross-task pool also includes M3AE and DC-Seg [14; 10], alongside the methods listed in the table.

Method	Year	Lesion	Gland	Avg.
RFNet [6]	2021	39.80	86.70	63.30
M3AE [14]	2023	40.60	86.90	63.80
DC-Seg [10]	2025	41.20	87.10	64.20
mmFormer [24]	2022	40.90	87.00	64.00
M^2FTrans [16]	2024	41.60	87.30	64.50
AdaMM [29]	2026	42.40	87.50	65.00
SimMLM [11]	2025	42.80	87.70	65.30
MoMKE [21]	2024	43.10	87.80	65.50
THEME w/o evidence	—	44.00	88.20	66.10
THEME	—	**46.80**	**88.90**	**67.90**

Table 4: Cross-task result on Prostate158, averaged over seven non-empty T2W/DWI/ADC subsets. Values are Dice (%).

Which sequence remains matters. The full matrices reveal target-specific behavior hidden by an all-subset mean. On BraTS 2020, the four single-sequence configurations span 81.8–90.8% WT Dice, 72.3–86.7% TC Dice, and 50.5–80.1% ET Dice for THEME. The larger ET range reflects the different information supplied by the individual sequences. Equal input count therefore does not imply equal segmentation difficulty. Across the one- and two-sequence BraTS 2020 configurations, mean ET Dice is 64.24%, compared with 76.00% for three- and four-sequence configurations. This 11.76-point difference quantifies the effect of sparse input for the inner tumor region.

Mean vector	R_1	R_2	R_3
R_1	1.000	0.069	0.146
R_2	0.069	1.000	0.238
R_3	0.146	0.238	1.000

Table 5: Cosine similarity of mean task-summary vectors.

The full-input BraTS 2020 endpoints exceed the corresponding all-subset means by 2.65/5.46/12.14 points for WT/TC/ET. Reporting only an average over the three targets would obscure this concentration of missing-input sensitivity in TC and ET. We therefore retain separate target columns and report a TC/ET summary for severe missingness. The qualitative T1+T2 cases show the same distinction spatially: a plausible broad WT outline can coexist with a missed or overexpanded inner region. Figure 2 presents the predicted masks next to the annotations so that these error patterns can be inspected directly.

4.5 REPRESENTATION AND COMPUTATIONAL COST

Table 5 reports cosine similarity between mean task summaries. The three off-diagonal values are 0.069, 0.146, and 0.238. The first task has lower mean-vector similarity to the others than the second and third tasks have to each other. Each query pools an encoder map separately, and each predicted summary is routed to its corresponding decoder task. The measured directions show that this task indexing persists through training: the three query outputs do not collapse to the same mean feature vector. The table measures representation geometry across validation cases; the segmentation tables measure whether these features improve the predicted regions.

The local inference audit uses one RTX 4090 and a $1 \times 4 \times 128^3$ patch. THEME has 8.59M parameters and 163.19 estimated GFLOPs. Compared with the same-expert model without the summary pathway, the additional modules require 0.74M parameters and 40.72 GFLOPs. Encoder experts account for 94.27% of parameters; the missing-summary predictor accounts for 4.90%, and task-spatial fusion accounts for 0.37%. The task pathway is thus a compact addition to the encoders. Its measured cost is 88.7 ms and 2194 MB, compared with 57.1 ms and 1031 MB for the same-expert variant. Table 6 reports the corresponding measurements for the other models under the same local patch protocol.

Method	Param.	GFLOPs	MB	ms
THEME	8.59	163.19	2194	88.7
THEME w/o evidence	7.85	122.47	1031	57.1
SimMLM	7.82	122.76	651	60.7
M^2FTrans	13.49	119.29	2114	410.3
mmFormer	36.65	130.19	1664	67.7
AdaMM	26.55	646.12	2267	67.8
RFNet	8.98	420.89	8012	163.7
M3AE	40.42	149.02	1138	26.1
DC-Seg	40.04	420.89	8136	164.9

Table 6: Inference efficiency under the unified local protocol detailed in the supplement. Parameters are in millions; GFLOPs are estimates, MB is peak allocated memory, and ms is latency.

5 CONCLUSION

THEME predicts task-conditioned summaries of missing MRI sequences and uses them to modulate a decoder whose spatial input comes from observed sequences. Modality–region indexing makes the completion target explicit, and separate spatial and task routing connects the acquired maps and the completed summaries to their respective roles in prediction. The framework learns this connection jointly through segmentation, summary reconstruction, and supervision across nested input sets.

Evaluation over all 15 non-empty modality subsets on BraTS 2018 and 2020 shows higher mean WT and ET Dice than the strongest aggregate comparator. The severe-missing TC/ET gains are 1.90 and 1.91 points on the two releases, respectively. Cumulative architecture ablations improve all three target scores on both datasets, while removing direct summary supervision produces the largest ET decrease among the tested loss groups. Prostate158 extends the same evaluation principle to a different organ and three input sequences, where the complete model improves on its no-evidence variant. Together, these results support learning missing information at the level of the region task that consumes it and combining that information with spatial features from the available examination. The per-subset results also show that equal numbers of available sequences can produce markedly different inner-region scores. This observation motivates reporting each availability configuration and each tumor target separately. In the model, a completed summary changes the region task's decoder context, while the spatial pathway continues to use feature maps extracted from the sequences supplied for that examination.

ETHICS STATEMENT

This work involved no direct interaction with human subjects, collection of personal information, or deployment decisions.

Reproducibility statement

The main paper and appendix document the method, training and evaluation protocols, and complete per-subset results. Anonymous code is available at https://anonymous.4open.science/r/THEME-D447.

AI use statement

Generative AI tools assisted with drafting, revising, and organizing the manuscript. The authors are responsible for the final text, analyses, and reported results.

References

[1] Danish Ali, Ajmal Mian, Naveed Akhtar, and Ghulam Mubashar Hassan. D3Seg: Dependency-Aware Diffusion for Brain Tumor Segmentation with Missing Modalities. arXiv preprint arXiv:2605.22249, 2026. URL https://arxiv.org/abs/2605.22249.

[2] Reza Azad, Nika Khosravi, and Dorit Merhof. SMU-Net: Style matching U-Net for brain tumor segmentation with missing modalities. arXiv preprint arXiv:2204.02961, 2022. URL https://arxiv.org/abs/2204.02961.

[3] Seunghun Baek, Jihwan Park, Jaeyoon Sim, Hoseok Lee, Seungjoo Lee, and Won Hwa Kim. Set-Inclusive Uncertainty Modeling for Robust Brain Tumor Segmentation. arXiv preprint arXiv:2606.30374, 2026. URL https://arxiv.org/abs/2606.30374.

[4] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, et al. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv preprint arXiv:1811.02629, 2018. URL https://arxiv.org/abs/1811.02629.

[5] Cheng Chen, Qi Dou, Yueming Jin, Hao Chen, Jing Qin, and Pheng-Ann Heng. Robust Multimodal Brain Tumor Segmentation via Feature Disentanglement and Gated Fusion. arXiv preprint arXiv:2002.09708, 2020. URL https://arxiv.org/abs/2002.09708.

[6] Yuhang Ding, Xin Yu, and Yi Yang. RFNet: Region-Aware Fusion Network for Incomplete Multi-Modal Brain Tumor Segmentation. In *Proceedings of ICCV*, pp. 3975–3984, 2021.

[7] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. HeMIS: Hetero-Modal Image Segmentation. arXiv preprint arXiv:1607.05194, 2016. URL https://arxiv.org/abs/1607.05194.

[8] Seungwan Jeong, Hwanho Cho, Junmo Kwon, and Hyunjin Park. Region-of-Interest Attentive Heteromodal Variational Encoder-Decoder for Segmentation with Missing Modalities. In *Proceedings of ACCV*, pp. 3707–3723, 2022.

[9] Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, and Prateek Prasanna. Enhancing Modality-Agnostic Representations via Meta-Learning for Brain Tumor Segmentation. arXiv preprint arXiv:2302.04308, 2023. URL https://arxiv.org/abs/2302.04308.

[10] Haitao Li, Ziyu Li, Yiheng Mao, Zhengyao Ding, and Zhengxing Huang. DC-Seg: Disentangled Contrastive Learning for Brain Tumor Segmentation with Missing Modalities. arXiv preprint arXiv:2505.11921, 2025. URL https://arxiv.org/abs/2505.11921.

[11] Sijie Li, Chen Chen, and Jungong Han. SimMLM: A Simple Framework for Multi-modal Learning with Missing Modality. arXiv preprint arXiv:2507.19264, 2025. URL https://arxiv.org/abs/2507.19264.

[12] 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. ISSN 1746-8094. doi: 10.1016/j.bspc.2024.106012. URL https://www.sciencedirect.com/science/article/pii/S1746809424000703.

[13] Guoyan Liang, Qin Zhou, Jingyuan Chen, Bingcang Huang, Kai Chen, Lin Gu, Zhe Wang, Sai Wu, and Chang Yao. Semantic-guided Masked Mutual Learning for Multi-modal Brain Tumor Segmentation with Arbitrary Missing Modalities. arXiv preprint arXiv:2507.07592, 2025. URL https://arxiv.org/abs/2507.07592.

[14] Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3AE: Multimodal Representation Learning for Brain Tumor Segmentation with Missing Modalities. arXiv preprint arXiv:2303.05302, 2023. URL https://arxiv.org/abs/2303.05302.

[15] Bjoern H. Menze, Andras Jakab, Stefan Bauer, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2015. doi: 10.1109/TMI.2014.2377694.

[16] Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. M^{2}FTrans: Modality-Masked Fusion Transformer for Incomplete Multi-Modality Brain Tumor Segmentation. *IEEE Journal of Biomedical and Health Informatics*, 28(1):379–390, 2024. doi: 10.1109/JBHI.2023.3326151.

[17] 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. doi: 10.1109/JBHI.2023.3286689.

[18] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal Learning with Missing Modality via Shared-Specific Feature Modelling. arXiv preprint arXiv:2307.14126, 2023. URL https://arxiv.org/abs/2307.14126.

[19] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable Cross-modal Knowledge Distillation for Multi-modal Learning with Missing Modality. arXiv preprint arXiv:2310.01035, 2023. URL https://arxiv.org/abs/2310.01035.

[20] Yixin Wang, Yang Zhang, Yang Liu, et al. ACN: Adversarial Co-training Network for Brain Tumor Segmentation with Missing Modalities. arXiv preprint arXiv:2106.14591, 2021. URL https://arxiv.org/abs/2106.14591.

[21] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging Knowledge of Modality Experts for Incomplete Multimodal Learning. In *ACM Multimedia 2024*, 2024. doi: 10.1145/3664647.3681683.

[22] 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. doi: 10.1109/TMI.2023.3301934.

[23] Ziqi Yu, Xiaoyang Han, Shengjie Zhang, Jianfeng Feng, Tingying Peng, and Xiao-Yong Zhang. MouseGAN++: Unsupervised Disentanglement and Contrastive Representation for Multiple MRI Modalities Synthesis and Structural Segmentation of Mouse Brain. arXiv preprint arXiv:2212.01825, 2022. URL https://arxiv.org/abs/2212.01825.

[24] Yao Zhang, Nanjun He, Jiawei Yang, et al. mmFormer: Multimodal Medical Transformer for Incomplete Multimodal Learning of Brain Tumor Segmentation. arXiv preprint arXiv:2206.02425, 2022. URL https://arxiv.org/abs/2206.02425.

[25] Zheyu Zhang, Gang Yang, Yueyi Zhang, Huanjing Yue, Aiping Liu, Yunwei Ou, Jian Gong, and Xiaoyan Sun. TMFormer: Token Merging Transformer for Brain Tumor Segmentation with Missing Modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7414–7422, 2024. doi: 10.1609/aaai.v38i7.28572.

[26] Xiaoqi Zhao, Youwei Pang, Chenyang Yu, et al. UniMRSeg: Unified Modality-Relax Segmentation via Hierarchical Self-Supervised Compensation. arXiv preprint arXiv:2509.16170, 2025. URL https://arxiv.org/abs/2509.16170.

[27] Shenghao Zhu, Yifei Chen, Shuo Jiang, et al. XLSTM-HVED: Cross-Modal Brain Tumor Segmentation and MRI Reconstruction Method Using Vision XLSTM and Heteromodal Variational Encoder-Decoder. arXiv preprint arXiv:2412.07804, 2024. URL https://arxiv.org/abs/2412.07804.

[28] Shenghao Zhu, Yifei Chen, Weihong Chen, et al. Bridging the Gap in Missing Modalities: Leveraging Knowledge Distillation and Style Matching for Brain Tumor Segmentation. arXiv preprint arXiv:2507.22626, 2025. URL https://arxiv.org/abs/2507.22626.

[29] Shenghao Zhu, Yifei Chen, Weihong Chen, et al. No Modality Left Behind: Adapting to Missing Modalities via Knowledge Distillation for Brain Tumor Segmentation. arXiv preprint arXiv:2509.15017, 2025. URL https://arxiv.org/abs/2509.15017.

Supplementary Material

This supplement provides implementation details, qualitative examples, and results for every non-empty modality subset.

Contents

A Method and implementation details 12
A.1 Loss terms 12
A.2 Data and optimization 13
B Qualitative case and full subset matrices 13
B.1 Per-subset Dice 13

A Method and implementation details

Each modality expert has a six-stage 3D encoder–decoder with feature widths $(8, 16, 32, 64, 80, 80)$ and independent parameters. Its final decoded map supplies the spatial branch, and its deepest encoder map supplies summary pooling. Each region uses two pooled tokens, aggregated by an MLP into a 64-dimensional summary. Extraction queries $p^{\mathrm{ext}}_{m,t}$, completion queries p^{cmp}_{t}, and routing queries p^{route}_{t} are separate parameters. The region masks used for extraction are resized by trilinear interpolation and clipped to $[0, 1]$. At inference they are constructed from detached expert probabilities; the region differences are clipped to $[0, 1]$ before resizing.

For global routing, let $a_m = \mathbf{1}[m \in \mathcal{A}]$ and let e_m be a learned modality embedding. The router predicts $u_m = \mathrm{softplus}(f_u(g_m))$ and computes $s_m = f_g([g_m; a_m; u_m; e_m]) - u_m$. The task score is decomposed as $w_{m,t} = s_m + \beta f_{\mathrm{task}}([\tilde{z}_{m,t}; p^{\mathrm{route}}_{t}; e_m; u_m]) - u_m$, with learned scalar β initialized to one. The global weights normalize over the acquired modalities. Task-summary aggregation is defined in Eq. 7.

In completion, the base b_t is shared across target modalities, while q_u identifies the modality to be predicted. The keys and values additionally receive a reliability scalar. With the completion call used in the available code, that scalar is one for every observed source and is a constant input to the projection MLPs. Intermediate estimates $\rho_{u,t}$ condition subsequent heads; the learned scales η_t are applied to the final residuals. The missing slots passed to fusion use $0.7b_t + 0.3\hat{z}_{u,t}$. Spatial modulation uses a 64-channel context projection. The learned γ_t values are initialized to 0.1, and the final gate convolution has zero weights and bias -2.

A.1 Loss terms

Let Y_t denote the resized mask for the three exclusive label regions and G_t the sigmoid modulation gate. The gate-alignment term in Eq. 12 uses task weights $(2, 1, 1)$. The ET mask is dilated on the gate grid by a $3 \times 3 \times 3$ max-pooling operation with stride one before alignment. Gate probabilities are clipped to $[10^{-4}, 1 - 10^{-4}]$ when computing BCE. The gate-area term is $\frac{1}{3}\sum_t \mathrm{mean}(G_t)$. For task-routing weights $\alpha_{m,t}$, the separation term is

$$\mathcal{L}_{\mathrm{route}} = \sum_{(s,t)\in\{(1,2),(1,3),(2,3)\}} \left[0.15 - \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} |\alpha_{m,s} - \alpha_{m,t}|\right]_+ . \tag{14}$$

The objective uses $\lambda_o = 0.1$, $\lambda_e = 1.0$, $\lambda_r = 0.05$, $\lambda_g = 0.05$, $\lambda_a = 0.005$, and $\lambda_d = 0.02$. Reconstruction compares the concatenated 192-dimensional modality summaries: the L1 term averages over their coordinates, and cosine distance is computed on the concatenated vector. Both terms average over hidden modality slots in the paired batch. Evidence ranking instead averages L1 error over all modality, task, and feature coordinates. Output ranking applies its hinge separately to Dice and BCE for each output region. Both partial views remain in the computation graph, and the full-input summaries are detached. The prototype-anchor regularizer is disabled.

A.2 Data and optimization

Inputs are normalized per sequence over non-zero foreground voxels and cropped to 128^3 patches. Training pads then randomly crops; validation center-crops then pads. All four sequences of a subject receive the same spatial transform. Training uses AdamW, initial/minimum learning rates $10^{-3}/10^{-6}$, a 300-epoch cosine schedule, weight decay 10^{-2}, and batch size two subject pairs. Each axis is flipped with probability 0.5. Noise, smoothing, contrast, and intensity scaling each have probability 0.15. Evidence supervision ramps over the first 20 epochs.

The complete test pool is $2^{\mathcal{M}} \setminus \{\varnothing\}$. It has four one-sequence sets, six two-sequence sets, four three-sequence sets, and one four-sequence set. The severe-missing score averages the first ten sets (one or two observed sequences). Subset means are unweighted. For each seed, the model produces one mean across the selected sets; the main tables summarize those five seed means. Full per-set matrices follow in Appendix B.1.

B Qualitative case and full subset matrices

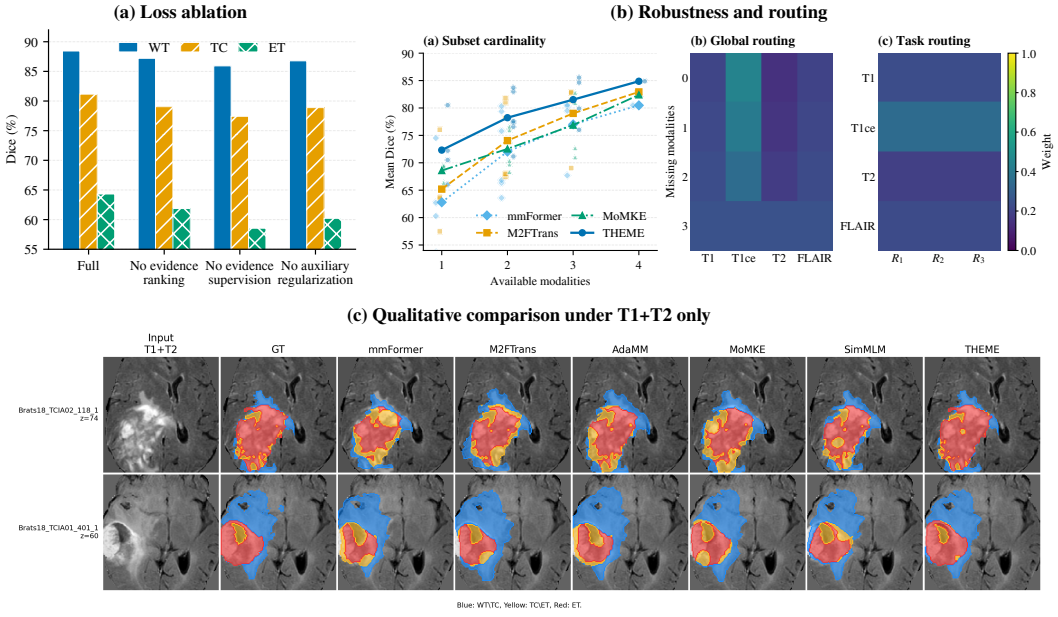

Figure S1: BraTS 2018 diagnostic figure. The qualitative cases use T1+T2 input, with both FLAIR and T1ce absent. The accompanying plots show loss ablations and input-cardinality behavior.

B.1 Per-subset Dice

The following matrices report each of the 15 non-empty input configurations for WT, TC, and ET on BraTS 2018 and 2020. Their mean and population standard deviation summarize variation across input sets; they are distinct from the five-run standard deviations in the main tables.

Index	Available	Index	Available
1	T2	9	FLAIR, T2
2	T1ce	10	FLAIR, T1ce
3	T1	11	FLAIR, T1, T1ce
4	FLAIR	12	FLAIR, T1, T2
5	T1ce, T2	13	FLAIR, T1ce, T2
6	T1, T1ce	14	T1, T1ce, T2
7	FLAIR, T1	15	FLAIR, T1, T1ce, T2
8	T1, T2		

Table S1: BraTS 2018 WT: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	76.18	71.32	70.72	74.31	75.28	74.93	76.73	77.26
ShaSpec [18]	2023	78.89	42.74	43.99	66.55	81.97	82.15	82.34	56.70
MriSeg [17]	2024	82.90	68.58	78.07	47.89	85.90	82.33	80.67	85.82
MouseGAN++ [23]	2023	66.67	75.83	60.28	60.51	79.07	79.71	71.36	70.63
Hyper-GAE [22]	2023	73.99	69.36	67.59	74.44	72.81	60.99	73.15	74.44
SRMNet [12]	2024	70.84	69.09	69.51	76.08	75.29	76.46	76.81	73.72
RFNet [6]	2021	68.28	55.71	56.04	71.45	69.62	60.33	73.62	71.32
mmFormer [24]	2022	81.13	73.15	69.75	87.15	85.26	76.74	88.25	83.21
M^2FTrans [16]	2024	83.49	74.67	73.21	87.45	86.16	78.64	88.28	85.06
RMBTS [5]	2019	78.76	79.57	67.98	61.35	77.87	76.11	72.55	73.75
RA-HVED [8]	2022	73.33	**87.00**	71.13	81.29	**89.42**	**87.30**	82.85	77.16
XLSTM-HVED [27]	2025	64.28	78.82	55.90	64.81	70.84	81.81	82.16	81.44
ACN [20]	2021	82.52	71.80	73.00	87.45	84.28	75.22	88.24	84.27
SMUNet [2]	2022	83.75	72.00	76.04	87.67	84.94	77.20	89.35	85.29
MSTKDNet [28]	2025	79.81	75.82	78.68	69.19	83.62	80.26	78.66	85.58
AdaMM [29]	2026	85.21	79.53	76.51	88.45	87.42	81.34	89.69	86.56
MoMKE [21]	2024	86.37	77.45	74.00	85.52	86.20	79.86	90.23	86.60
SimMLM [11]	2025	86.79	78.30	75.23	86.22	87.00	80.03	90.30	85.97
THEME	—	**87.14**	80.34	**81.23**	**89.60**	87.45	86.91	**90.56**	**87.12**

Table S2: BraTS 2018 WT: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	78.31	79.16	79.94	80.59	81.08	81.12	81.56	77.23	3.35
ShaSpec [18]	2023	69.91	71.98	83.19	83.53	83.75	73.32	84.41	72.36	13.70
MriSeg [17]	2024	85.60	66.20	80.71	86.66	87.86	87.85	87.69	79.65	10.63
MouseGAN++ [23]	2023	68.60	79.88	79.10	73.00	81.20	80.14	82.81	73.92	7.13
Hyper-GAE [22]	2023	76.17	75.43	78.80	78.27	79.24	77.75	80.62	74.20	4.95
SRMNet [12]	2024	77.53	78.60	78.67	77.66	78.76	80.86	84.73	76.31	4.07
RFNet [6]	2021	75.24	73.48	73.99	75.94	75.77	70.72	75.92	69.83	6.69
mmFormer [24]	2022	88.19	88.73	88.71	88.43	89.18	85.81	89.12	84.19	6.05
M^2FTrans [16]	2024	88.30	88.46	88.81	88.58	88.93	86.40	89.12	85.04	5.11
RMBTS [5]	2019	72.19	79.28	80.14	84.40	82.92	80.59	83.56	76.73	6.06
RA-HVED [8]	2022	84.02	87.55	87.85	84.26	89.23	89.31	88.98	84.05	5.73
XLSTM-HVED [27]	2025	66.84	72.80	83.04	83.38	83.46	73.93	83.74	75.15	8.60
ACN [20]	2021	87.60	87.47	88.21	88.34	87.67	84.37	88.29	83.92	5.61
SMUNet [2]	2022	88.25	88.14	89.26	89.39	88.31	85.59	89.26	84.96	5.32
MSTKDNet [28]	2025	75.51	77.05	79.38	83.99	79.73	85.81	84.84	79.86	4.35
AdaMM [29]	2026	89.47	**90.59**	90.56	89.63	90.73	87.74	90.62	86.94	4.30
MoMKE [21]	2024	90.53	90.23	90.21	89.03	90.35	86.42	90.34	86.22	5.01
SimMLM [11]	2025	88.13	89.23	90.63	88.11	90.34	87.08	90.83	86.28	4.58
THEME	—	**90.78**	90.43	**90.77**	**91.14**	**91.09**	**90.77**	**91.25**	**88.44**	3.39

Table S3: BraTS 2018 TC: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	61.97	60.33	66.64	66.16	66.23	68.71	70.65	72.02
ShaSpec [18]	2023	73.95	45.03	66.21	65.42	76.53	74.90	75.83	75.31
MriSeg [17]	2024	59.39	75.37	61.07	32.76	79.48	79.84	61.83	61.69
MouseGAN++ [23]	2023	61.92	67.87	60.25	61.79	68.47	60.86	66.28	67.02
Hyper-GAE [22]	2023	57.42	62.28	54.38	61.06	72.97	67.42	60.48	61.06
SRMNet [12]	2024	62.90	74.62	54.96	61.63	71.95	80.55	67.76	66.03
RFNet [6]	2021	53.33	65.47	46.65	55.54	72.07	68.74	59.62	57.83
mmFormer [24]	2022	62.76	81.63	59.72	64.00	82.48	81.14	71.64	68.10
M^2FTrans [16]	2024	67.71	82.25	64.82	67.47	83.75	83.46	72.60	72.88
RMBTS [5]	2019	61.17	56.77	75.69	44.45	68.19	82.37	78.28	81.92
RA-HVED [8]	2022	**80.57**	64.12	60.31	67.11	81.95	67.42	70.23	81.81
XLSTM-HVED [27]	2025	17.08	18.65	50.41	25.89	52.45	55.10	31.36	23.52
ACN [20]	2021	69.28	62.80	82.03	68.44	68.69	81.13	83.38	83.24
SMUNet [2]	2022	68.98	61.96	81.68	69.91	70.28	82.89	**84.25**	84.02
MSTKDNet [28]	2025	69.28	68.32	**83.94**	55.34	73.44	84.18	81.92	84.76
AdaMM [29]	2026	73.35	**85.79**	70.98	74.18	**86.56**	**85.92**	75.99	74.68
MoMKE [21]	2024	72.34	78.66	68.71	**74.34**	83.69	84.62	75.00	73.73
SimMLM [11]	2025	73.12	83.45	71.48	74.12	84.19	84.79	77.14	74.73
THEME	—	72.58	84.90	70.18	74.23	83.43	85.69	72.98	**86.48**

Table S4: BraTS 2018 TC: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	71.74	71.62	72.73	73.54	73.37	74.01	74.64	69.62	4.30
ShaSpec [18]	2023	65.60	77.48	80.20	77.12	81.08	79.61	82.79	73.14	9.18
MriSeg [17]	2024	62.82	74.85	80.16	66.67	79.90	80.43	81.16	69.16	12.76
MouseGAN++ [23]	2023	67.71	70.33	72.58	67.85	70.97	73.02	74.75	67.44	4.41
Hyper-GAE [22]	2023	63.09	73.51	79.76	72.58	76.61	80.63	82.42	68.38	8.80
SRMNet [12]	2024	67.23	71.79	83.20	68.37	72.64	82.65	83.14	71.29	8.19
RFNet [6]	2021	59.84	73.85	73.90	61.95	75.67	73.08	75.67	64.88	8.92
mmFormer [24]	2022	69.74	82.58	82.48	72.34	82.66	82.09	82.51	75.06	8.22
M^2FTrans [16]	2024	72.67	84.41	84.75	74.26	84.67	84.29	84.67	77.64	7.22
RMBTS [5]	2019	65.51	63.21	82.71	82.19	68.70	82.83	82.99	71.80	11.45
RA-HVED [8]	2022	82.20	68.61	70.81	82.99	82.58	82.82	82.93	75.10	7.99
XLSTM-HVED [27]	2025	55.04	27.45	31.27	60.09	61.50	62.10	62.72	42.31	16.87
ACN [20]	2021	69.69	69.46	83.32	83.65	70.64	83.11	82.83	76.11	7.39
SMUNet [2]	2022	71.35	70.95	84.34	**84.22**	71.69	84.12	84.07	76.98	7.52
MSTKDNet [28]	2025	61.68	62.83	81.78	83.69	66.95	85.00	83.78	75.13	9.87
AdaMM [29]	2026	75.57	**86.42**	**86.35**	76.47	86.59	86.41	**86.94**	80.81	6.07
MoMKE [21]	2024	75.56	84.62	82.25	77.02	83.23	82.35	82.61	78.58	4.95
SimMLM [11]	2025	76.90	84.96	85.25	78.02	85.05	85.08	85.39	80.24	5.09
THEME	—	**85.32**	84.51	72.47	83.79	**87.01**	**86.85**	**86.94**	**81.16**	6.26

Table S5: BraTS 2018 ET: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	38.36	35.38	46.97	44.73	43.91	48.44	51.73	54.20
ShaSpec [18]	2023	30.97	21.24	46.20	25.80	31.29	62.90	53.01	56.71
MriSeg [17]	2024	31.34	66.58	35.25	14.88	68.93	68.61	36.30	37.33
MouseGAN++ [23]	2023	38.33	41.27	27.28	36.08	41.72	51.38	40.72	41.03
Hyper-GAE [22]	2023	34.52	30.72	36.55	38.12	36.79	54.52	48.20	53.11
SRMNet [12]	2024	35.19	44.55	26.96	33.24	60.80	58.61	35.91	36.54
RFNet [6]	2021	28.13	56.69	17.30	26.53	61.93	59.78	29.44	32.29
mmFormer [24]	2022	36.97	68.75	31.48	37.02	70.37	69.46	39.14	39.47
M^2FTrans [16]	2024	40.13	71.18	34.43	35.55	**73.19**	72.37	41.36	45.63
RMBTS [5]	2019	34.66	33.71	46.38	17.35	40.85	50.41	48.31	49.75
RA-HVED [8]	2022	39.28	45.62	56.18	46.11	36.65	43.58	43.80	38.44
XLSTM-HVED [27]	2025	13.94	14.93	50.98	21.82	53.08	56.37	28.33	19.66
ACN [20]	2021	42.25	30.24	69.84	40.27	41.74	70.06	72.06	72.70
SMUNet [2]	2022	42.73	34.21	71.21	41.49	42.22	72.38	72.98	74.04
MSTKDNet [28]	2025	32.61	40.67	**73.82**	35.48	41.71	73.09	**73.53**	**74.99**
AdaMM [29]	2026	49.69	**76.92**	43.28	46.52	56.71	**76.87**	48.18	50.05
MoMKE [21]	2024	48.16	52.66	65.47	39.33	52.38	44.75	62.74	44.74
SimMLM [11]	2025	49.15	49.79	62.86	45.47	54.26	45.46	68.89	48.09
THEME	—	**56.92**	76.28	46.70	**47.69**	58.87	76.38	49.93	58.35

Table S6: BraTS 2018 ET: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	52.96	51.89	53.71	55.18	54.35	55.59	56.65	49.60	6.28
ShaSpec [18]	2023	45.44	62.57	63.05	33.34	61.23	64.26	66.37	48.29	15.36
MriSeg [17]	2024	34.95	67.35	70.37	39.10	68.99	69.67	70.34	52.00	18.78
MouseGAN++ [23]	2023	48.91	50.10	58.22	43.25	58.53	56.65	59.46	46.20	9.18
Hyper-GAE [22]	2023	40.47	52.74	53.36	38.81	53.26	57.50	60.47	45.94	9.33
SRMNet [12]	2024	39.95	50.55	61.81	38.57	63.15	62.63	63.89	47.49	12.74
RFNet [6]	2021	33.60	62.32	62.48	35.35	64.47	62.87	64.86	46.54	16.98
mmFormer [24]	2022	42.57	69.69	69.81	42.29	69.77	70.59	69.82	55.15	15.83
M^2FTrans [16]	2024	43.03	72.60	74.85	44.27	74.91	73.46	75.00	58.13	16.62
RMBTS [5]	2019	37.39	34.82	51.08	59.82	40.74	65.39	64.38	45.00	12.41
RA-HVED [8]	2022	39.05	44.42	43.49	36.35	36.60	36.24	36.23	41.47	5.33
XLSTM-HVED [27]	2025	**56.81**	24.07	28.25	62.11	64.20	64.23	65.55	41.62	19.49
ACN [20]	2021	42.53	38.42	71.21	71.94	42.75	73.69	71.51	56.75	16.18
SMUNet [2]	2022	43.59	42.05	72.94	72.53	43.35	72.89	72.49	58.07	15.76
MSTKDNet [28]	2025	28.72	40.93	73.28	**73.85**	36.48	75.26	74.07	56.57	18.88
AdaMM [29]	2026	51.03	**77.66**	75.81	49.92	77.52	**77.23**	**79.30**	62.45	14.19
MoMKE [21]	2024	45.10	55.08	75.97	47.08	44.56	74.05	74.24	55.09	11.91
SimMLM [11]	2025	50.45	55.65	75.02	49.01	48.30	74.33	73.80	56.70	10.75
THEME	—	56.71	76.18	**76.40**	53.13	**78.51**	76.31	76.44	**64.32**	12.01

Table S7: BraTS 2020 WT: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	77.5	72.8	72.1	75.8	76.8	76.3	78.1	78.6
ShaSpec [18]	2023	80.1	45.2	46.4	68.1	83.2	83.4	83.5	58.9
MriSeg [17]	2024	84.2	70.3	79.4	50.1	87.1	83.6	81.9	86.9
MouseGAN++ [23]	2023	68.2	77.1	62.1	62.3	80.4	81.0	72.8	72.1
Hyper-GAE [22]	2023	75.3	70.9	69.2	75.9	74.2	62.8	74.6	75.8
SRMNet [12]	2024	72.3	70.6	71.1	77.5	76.8	77.9	78.2	75.2
RFNet [6]	2021	69.8	57.6	58.0	72.9	71.2	62.3	75.1	72.8
mmFormer [24]	2022	82.5	74.8	71.5	88.3	86.6	78.3	89.4	84.6
M^2FTrans [16]	2024	84.8	76.2	74.8	88.6	87.4	80.2	89.5	86.4
RMBTS [5]	2019	80.2	81.0	69.7	63.4	79.3	77.6	74.1	75.2
RA-HVED [8]	2022	74.9	88.2	72.8	82.7	90.5	88.4	84.2	78.7
XLSTM-HVED [27]	2025	65.9	80.2	57.8	66.5	72.4	83.1	83.5	82.8
ACN [20]	2021	83.9	73.5	74.7	88.6	85.6	76.9	89.4	85.6
SMUNet [2]	2022	85.1	73.7	77.6	88.8	86.3	78.8	90.5	86.6
MSTKDNet [28]	2025	81.2	77.4	80.2	70.9	84.9	81.7	80.2	86.8
AdaMM [29]	2026	86.5	81.1	78.2	**89.6**	88.6	82.8	90.8	87.8
MoMKE [21]	2024	87.6	79.1	75.8	86.8	87.5	81.4	91.3	87.8
SimMLM [11]	2025	88.0	79.9	77.0	87.4	88.2	81.6	**91.4**	87.2
THEME	—	**90.8**	**88.5**	**81.8**	82.6	**91.9**	**88.8**	88.2	**91.6**

Table S8: BraTS 2020 WT: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	79.7	80.4	81.1	81.8	82.4	82.5	82.8	78.58	3.29
ShaSpec [18]	2023	71.3	73.4	84.4	84.7	84.9	74.8	85.6	73.86	13.26
MriSeg [17]	2024	86.8	68.1	82.0	87.8	88.9	88.9	88.7	80.98	10.31
MouseGAN++ [23]	2023	70.1	81.2	80.4	74.6	82.5	81.4	84.1	75.35	6.96
Hyper-GAE [22]	2023	77.5	76.8	80.1	79.5	80.6	79.1	81.8	75.61	4.80
SRMNet [12]	2024	78.9	80.0	80.1	79.1	80.2	82.2	85.9	77.73	3.99
RFNet [6]	2021	76.7	74.9	75.4	77.3	77.2	72.3	77.4	71.39	6.51
mmFormer [24]	2022	89.4	89.9	89.8	89.6	90.3	87.1	90.3	85.49	5.86
M^2FTrans [16]	2024	89.6	89.7	90.0	89.8	90.1	87.7	90.3	86.34	4.97
RMBTS [5]	2019	73.7	80.7	81.5	85.6	84.3	82.1	84.9	78.22	5.89
RA-HVED [8]	2022	85.4	88.7	89.0	85.6	90.3	90.4	90.1	85.33	5.54
XLSTM-HVED [27]	2025	68.5	74.4	84.4	84.6	84.7	75.5	85.0	76.62	8.41
ACN [20]	2021	88.8	88.7	89.4	89.5	88.9	85.7	89.5	85.25	5.42
SMUNet [2]	2022	89.4	89.3	90.4	90.5	89.5	86.9	90.4	86.25	5.14
MSTKDNet [28]	2025	77.1	78.6	80.8	85.3	81.2	87.1	86.1	81.30	4.22
AdaMM [29]	2026	90.6	**91.6**	91.6	90.8	91.8	88.9	91.7	88.16	4.11
MoMKE [21]	2024	91.6	91.3	91.2	90.2	91.4	87.7	91.4	87.47	4.78
SimMLM [11]	2025	89.3	90.3	91.7	89.3	91.4	88.3	91.9	87.53	4.37
THEME	—	**91.7**	88.5	**91.9**	**91.9**	**92.2**	**92.1**	**92.3**	**89.65**	3.27

Table S9: BraTS 2020 TC: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	63.8	62.3	68.5	68.1	68.2	70.6	72.5	73.9
ShaSpec [18]	2023	75.8	47.4	68.2	67.4	78.4	76.8	77.7	77.2
MriSeg [17]	2024	61.5	77.3	63.2	35.1	81.3	81.7	63.9	63.8
MouseGAN++ [23]	2023	63.9	69.8	62.4	63.8	70.4	62.9	68.3	69.0
Hyper-GAE [22]	2023	59.5	64.3	56.6	63.1	74.9	69.4	62.6	63.1
SRMNet [12]	2024	64.9	76.5	57.2	63.7	73.9	82.4	69.8	68.1
RFNet [6]	2021	55.5	67.5	48.9	57.7	74.0	70.7	61.8	60.0
mmFormer [24]	2022	64.8	83.4	61.9	66.2	84.3	82.9	73.6	70.2
M^2FTrans [16]	2024	69.8	84.0	67.0	69.6	85.5	85.2	74.6	74.9
RMBTS [5]	2019	63.2	58.9	77.6	46.8	70.2	84.1	80.2	83.7
RA-HVED [8]	2022	**82.4**	66.3	62.5	69.2	83.8	69.5	72.3	83.6
XLSTM-HVED [27]	2025	19.5	21.0	52.6	28.2	54.7	57.3	33.7	25.9
ACN [20]	2021	71.4	64.9	83.8	70.5	70.7	82.9	85.1	85.0
SMUNet [2]	2022	71.1	64.1	83.5	72.0	72.3	84.6	86.0	85.8
MSTKDNet [28]	2025	71.4	70.4	85.7	57.6	75.5	85.9	83.7	**86.5**
AdaMM [29]	2026	75.5	**87.5**	73.1	76.3	**88.3**	**87.6**	78.0	76.7
MoMKE [21]	2024	74.5	80.6	70.8	**76.5**	85.5	86.4	77.0	75.8
SimMLM [11]	2025	75.3	85.2	73.6	76.2	85.9	86.5	79.2	76.8
THEME	—	76.4	74.7	**86.7**	72.3	87.1	85.2	**87.5**	86.3

Table S10: BraTS 2020 TC: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	73.6	73.5	74.6	75.4	75.2	75.8	76.4	71.49	4.26
ShaSpec [18]	2023	67.6	79.3	82.0	79.0	82.8	81.4	84.5	75.03	9.03
MriSeg [17]	2024	64.9	76.8	81.9	68.7	81.7	82.2	82.9	71.13	12.59
MouseGAN++ [23]	2023	69.7	72.3	74.5	69.8	72.9	74.9	76.6	69.41	4.35
Hyper-GAE [22]	2023	65.2	75.4	81.6	74.5	78.5	82.4	84.2	70.35	8.68
SRMNet [12]	2024	69.3	73.7	84.9	70.4	74.6	84.4	84.9	73.25	8.05
RFNet [6]	2021	61.9	75.8	75.8	64.1	77.6	74.9	77.5	66.91	8.79
mmFormer [24]	2022	71.8	84.3	84.2	74.4	84.4	83.8	84.2	76.96	8.05
M^2FTrans [16]	2024	74.7	86.1	86.4	76.3	86.4	86.0	86.3	79.52	7.04
RMBTS [5]	2019	67.6	65.3	84.5	84.0	70.7	84.6	84.7	73.74	11.28
RA-HVED [8]	2022	**84.0**	70.7	72.9	84.8	84.4	84.6	84.7	77.05	7.84
XLSTM-HVED [27]	2025	57.2	29.8	33.6	62.3	63.7	64.2	64.8	44.57	16.77
ACN [20]	2021	71.8	71.5	85.1	85.4	72.7	84.9	84.5	78.01	7.24
SMUNet [2]	2022	73.4	73.0	86.1	**86.0**	73.8	85.9	85.8	78.89	7.36
MSTKDNet [28]	2025	63.8	64.9	83.6	85.5	69.1	86.8	85.5	77.06	9.70
AdaMM [29]	2026	77.6	88.1	88.0	78.5	**88.3**	88.1	**88.6**	82.68	5.87
MoMKE [21]	2024	77.6	86.4	84.1	79.1	85.1	84.2	84.4	80.53	4.82
SimMLM [11]	2025	78.9	86.7	87.0	80.1	86.8	86.8	87.1	82.14	4.91
THEME	—	75.1	**88.2**	**88.6**	74.6	85.6	**88.8**	88.5	**83.04**	6.09

Table S11: BraTS 2020 ET: subsets 1–8. Bold marks a displayed column maximum.

Method	Year	1	2	3	4	5	6	7	8
FLAIR		○	○	○	●	○	○	●	○
T1		○	○	●	○	○	●	●	●
T1ce		○	●	○	○	●	●	○	○
T2		●	○	○	○	●	○	○	●
LCKD [19]	2023	42.1	39.2	50.8	48.6	47.7	52.3	55.6	58.1
ShaSpec [18]	2023	34.8	25.1	50.1	29.7	35.2	66.8	56.9	60.6
MriSeg [17]	2024	35.2	70.4	39.1	18.7	72.8	72.5	40.2	41.2
MouseGAN++ [23]	2023	42.2	45.1	31.1	39.9	45.6	55.2	44.6	44.9
Hyper-GAE [22]	2023	38.4	34.6	40.4	42.0	40.6	58.4	52.1	57.0
SRMNet [12]	2024	39.0	48.4	30.8	37.1	64.7	62.5	39.8	40.4
RFNet [6]	2021	32.0	60.5	21.2	30.4	65.8	63.6	33.3	36.1
mmFormer [24]	2022	40.8	72.6	35.3	40.9	74.2	73.3	43.0	43.3
M^2FTrans [16]	2024	44.0	75.0	38.3	39.4	**77.0**	76.2	45.2	49.5
RMBTS [5]	2019	38.5	37.6	50.2	21.2	44.7	54.3	52.2	53.6
RA-HVED [8]	2022	43.1	49.5	60.0	50.0	40.5	47.4	47.7	42.3
XLSTM-HVED [27]	2025	17.8	18.8	54.8	25.7	56.9	60.2	32.2	23.5
ACN [20]	2021	46.1	34.1	73.7	44.1	45.6	73.9	75.9	76.6
SMUNet [2]	2022	46.6	38.1	75.1	45.3	46.1	76.2	76.8	77.9
MSTKDNet [28]	2025	36.5	44.5	77.7	39.3	45.6	76.9	77.4	78.8
AdaMM [29]	2026	**53.5**	**80.8**	47.1	50.4	60.6	**80.7**	52.0	53.9
MoMKE [21]	2024	52.0	56.5	69.3	43.2	56.2	48.6	66.6	48.6
SimMLM [11]	2025	53.0	53.6	66.7	49.3	58.1	49.3	72.7	51.9
THEME	—	51.5	60.8	**80.1**	**50.5**	60.6	62.7	**80.2**	**80.0**

Table S12: BraTS 2020 ET: subsets 9–15 and across-subset mean and standard deviation. Bold marks a displayed column maximum.

Method	Year	9	10	11	12	13	14	15	Avg.	Std.
FLAIR		●	●	●	●	●	○	●		
T1		○	○	●	●	○	●	●		
T1ce		○	●	●	○	●	●	●		
T2		●	○	○	●	●	●	●		
LCKD [19]	2023	56.8	55.7	57.6	59.1	58.2	59.4	60.5	53.45	6.30
ShaSpec [18]	2023	49.3	66.4	66.9	37.2	65.1	68.1	70.2	52.16	15.36
MriSeg [17]	2024	38.8	71.2	74.2	43.0	72.8	73.5	74.2	55.85	18.77
MouseGAN++ [23]	2023	52.8	54.0	62.1	47.1	62.4	60.5	63.3	50.05	9.18
Hyper-GAE [22]	2023	44.3	56.6	57.2	42.7	57.1	61.4	64.3	49.81	9.33
SRMNet [12]	2024	43.8	54.4	65.7	42.4	67.0	66.5	67.7	51.35	12.75
RFNet [6]	2021	37.4	66.2	66.3	39.2	68.3	66.7	68.7	50.38	16.97
mmFormer [24]	2022	46.4	73.5	73.7	46.1	73.6	74.4	73.7	58.99	15.84
M^2FTrans [16]	2024	46.9	76.4	78.7	48.1	78.8	77.3	78.9	61.98	16.61
RMBTS [5]	2019	41.2	38.7	54.9	63.7	44.6	69.2	68.2	48.85	12.40
RA-HVED [8]	2022	42.9	48.3	47.3	40.2	40.5	40.1	40.1	45.33	5.32
XLSTM-HVED [27]	2025	**60.7**	27.9	32.1	65.9	68.1	68.1	69.4	45.47	19.48
ACN [20]	2021	46.4	42.3	75.1	75.8	46.6	77.5	75.4	60.61	16.18
SMUNet [2]	2022	47.4	45.9	76.8	76.4	47.2	76.7	76.3	61.92	15.76
MSTKDNet [28]	2025	32.6	44.8	77.1	77.7	40.3	79.1	77.9	60.41	18.87
AdaMM [29]	2026	54.9	**81.5**	79.6	53.8	**81.4**	81.1	**83.1**	66.29	14.18
MoMKE [21]	2024	48.9	58.9	79.8	50.9	48.4	77.9	78.1	58.93	11.91
SimMLM [11]	2025	54.3	59.5	78.9	52.9	52.1	78.2	77.6	60.54	10.75
THEME	—	53.8	62.2	**80.1**	**80.2**	57.0	**82.4**	80.3	**68.16**	12.01