Complementary Information Mutual Learning for Multimodality Medical Image Segmentation

Chuyun Shen   CYSHEN@STU.ECNU.EDU.CN
School of Computer Science and Technology
East China Normal University
Shanghai 200062, China

Wenhao Li   LIWENHAO@CUHK.EDU.CN
School of Data Science
The Chinese University of Hong Kong, Shenzhen
Shenzhen Institute of Artificial Intelligence and Robotics for Society
Shenzhen 518172, China

Haoqing Chen   51215901005@STU.ECNU.EDU.CN
School of Computer Science and Technology
East China Normal University
Shanghai 200062, China

Xiaoling Wang   XLWANG@CS.ECNU.EDU.CN
School of Computer Science and Technology
East China Normal University
Shanghai 200062, China

Fengping Zhu   ZHUFENGPING@FUDAN.EDU.CN
Huashan Hospital
Fudan University
Shanghai 200040, China

Yuxin Li   LIYUXIN@FUDAN.EDU.CN
Huashan Hospital
Fudan University
Shanghai 200040, China

Xiangfeng Wang   XFWANG@CS.ECNU.EDU.CN
School of Computer Science and Technology
East China Normal University
Shanghai AI Laboratory
Shanghai 200062, China

Bo Jin   BJIN@TONGJI.EDU.CN
School of Software Engineering
Shanghai Research Institute for Intelligent Autonomous Systems
Tongji University
Shanghai 200092, China

Abstract

Radiologists must utilize medical images of multiple modalities for tumor segmentation and diagnosis due to the limitations of medical imaging technology and the diversity of tumor signals. This has led to the development of multimodal learning in medical image

segmentation. However, the redundancy among modalities creates challenges for existing *subtraction*-based joint learning methods, such as misjudging the importance of modalities, ignoring specific modal information, and increasing cognitive load. These thorny issues ultimately decrease segmentation accuracy and increase the risk of overfitting. This paper presents the **complementary information mutual learning (CIML)** framework, which can mathematically model and address the negative impact of inter-modal redundant information. CIML adopts the idea of *addition* and removes inter-modal redundant information through inductive bias-driven task decomposition and message passing-based redundancy filtering. CIML first decomposes the multimodal segmentation task into multiple subtasks based on expert prior knowledge, minimizing the information dependence between modalities. Furthermore, CIML introduces a scheme in which each modality can extract information from other modalities additively through message passing. To achieve non-redundancy of extracted information, the redundant filtering is transformed into complementary information learning inspired by the variational information bottleneck. The complementary information learning procedure can be efficiently solved by variational inference and cross-modal spatial attention. Numerical results from the verification task and standard benchmarks indicate that CIML efficiently removes redundant information between modalities, outperforming SOTA methods regarding validation accuracy and segmentation effect. To emphasize, message-passing-based redundancy filtering allows neural network visualization techniques to visualize the knowledge relationship among different modalities, which reflects interpretability.

1. Introduction

The capacity of humans to develop a refined comprehension of their external environment can be attributed to the synergistic effects of multiple correlated sensory stimuli. This interaction results in emergent complexity, where the entirety of the system surpasses the simple sum of its individual components (Cohen et al., 1997; Gazzaniga et al., 2006). In the field of neuroscience, attention[1] focused on modality-rich spatiotemporal events optimizes information acquisition (Li et al., 2008; Li, 2023). Consequently, this enables individuals to utilize multimodal data for enhanced informational gain.

The significance of multimodal information in influencing human cognition and its implications for the development of artificial intelligence (AI) systems that emulate human intelligence has been widely recognized (McCarthy et al., 2006). This recognition has led to the advancement of multimodal learning (MML) as a key research area in AI (Baltrušaitis et al., 2018). MML represents a general framework for constructing AI models capable of extracting and relating information from multimodal data (Xu et al., 2022). In recent years, significant advances have been made in MML, particularly in computer vision and natural language processing (Bayoudh et al., 2021). Large-scale MML models have achieved near-human performance on specific tasks (Ramesh et al., 2021; Reed et al., 2022; Rombach et al., 2022).

This paper focuses on medical image segmentation, an essential task in medical image analysis that involves assigning labels to each pixel or voxel to identify distinct organs, tissues, or lesions. The intricate pathological or physiological features of human tissues

---

1. Upon encountering external stimuli, humans possess the ability to selectively concentrate on specific elements within these stimuli. This attention mechanism constitutes a fundamental aspect of human cognitive capacity, augmenting the efficacy of information processing (Zhang et al., 2012).
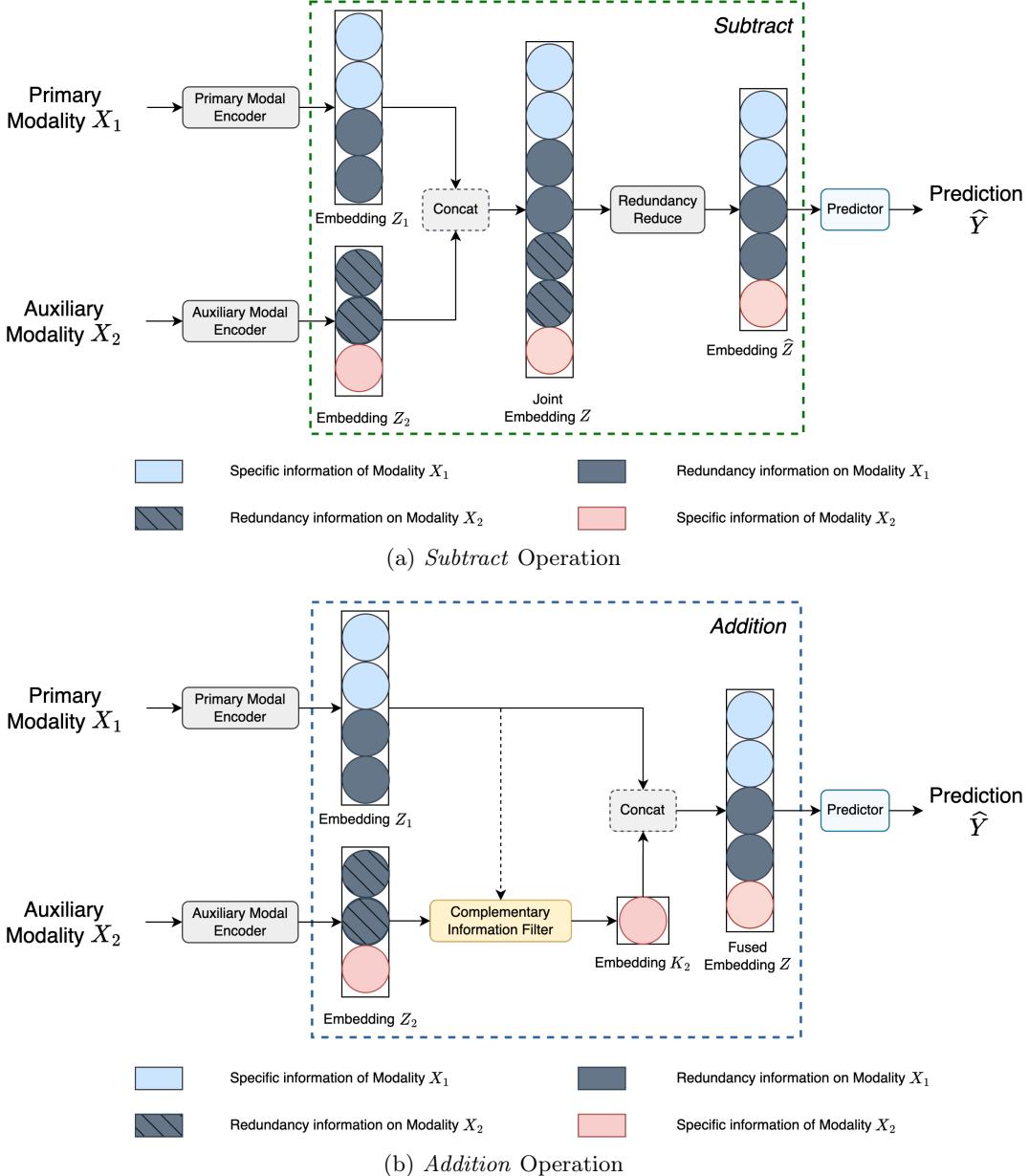
(a) *Subtract* Operation

(b) *Addition* Operation

Figure 1: The Diagram of *Addition* and *Subtract* operations. The dark gray circle with black slashes and the dark gray circle without black slashes represent the same information in embeddings. The *Subtract* operation first concatenates the information from different modalities and eliminates redundancy. The *Addition* operation first eliminates cross-modal redundant information and then concatenates the embeddings.

and lesions, combined with the variable sensitivity of imaging technologies across different human body components, necessitate the use of multimodal medical images for patient diagnosis and treatment. For instance, clinical guidelines for spontaneous intracerebral hemorrhage indicate that determining appropriate strategies relies on the mismatch between

two magnetic resonance imaging modalities: perfusion and diffusion imaging (Greenberg et al., 2022). Consequently, MML has become increasingly prevalent in medical image segmentation (Zhou et al., 2019).

Different from the modal-intensive events experienced by humans, machine learning problems with raw multimodal often present unaligned data (Wei et al., 2023), highlighting *modality alignment* and *modality synergy* as two thorny issues in MML. Fortunately, given the advanced understanding of the human body in medicine, image registration techniques (Hill et al., 2001) have matured, enabling the alignment of different medical modalities. Multimodal medical image segmentation emphasizes the *modality synergy* problem, namely constructing knowledge relationships among modalities (Han et al., 2022) to enable better information complementation and fusion.
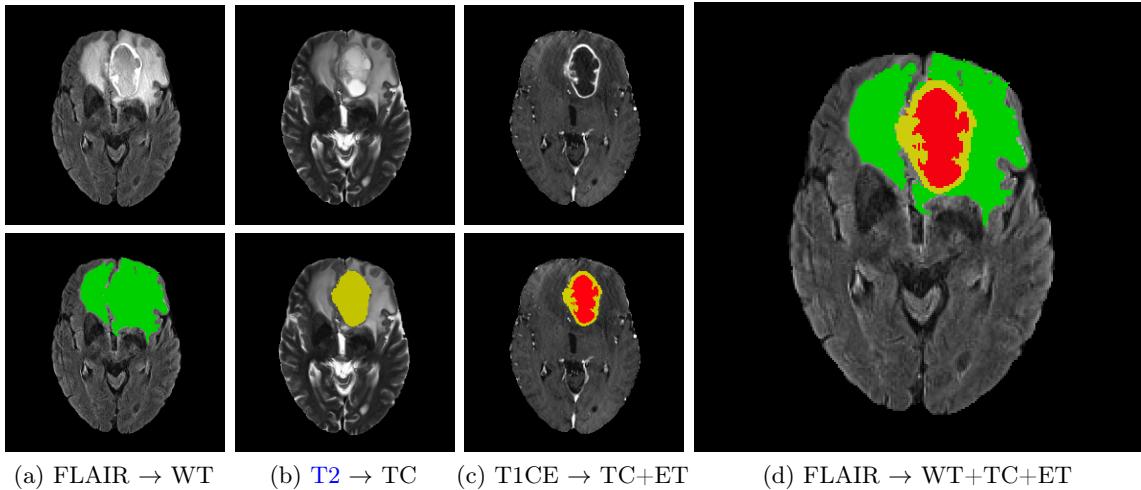

(a) FLAIR → WT   (b) T2 → TC   (c) T1CE → TC+ET   (d) FLAIR → WT+TC+ET

Figure 2: Dataset annotation for `BraTS2020`. Displayed are image patches with tumor structures annotated in various modalities (bottom left) and the final labels for the entire dataset (right). Image patches show from left to right: (a) the FLAIR image and the whole tumor (WT) visible in FLAIR; (b) the T2 image and the tumor core (TC) in T2; (c) the T1CE image, and the enhancing tumor (ET) visible in T1CE (yellow), surrounding the necrotic and non-enhancing tumor core(red); (d) Final labels of the tumor structures: edema (green), ET (yellow), necrotic and non-enhancing tumor core (red). In the `BraTS2020`challenge, images are requested to segment into WT, TC, and ET regions.

Multimodal medical image segmentation approaches are commonly designed with an end-to-end scheme to learn intermodal associations (Isensee et al., 2018; Zhang et al., 2021a,b; Zhou et al., 2020a, 2022; Ding et al., 2021; Dolz et al., 2018). Certain medical conditions (Figure 2) require segmentor to simultaneously identify multiple regions, such as the tumors, edemas, and necrotic tumor cores. We then refer to the end-to-end learning scheme above as *joint learning*, as it jointly maps multimodal inputs to single or multiple regions. The joint learning method typically involves fusing multimodal image encoding into a deep encoder-decoder architecture, which outputs the segmentation(s). These methods can be broadly classified into early fusion and mid-term fusion. The former directly concatenates

multimodal images as input to the network (Oktay et al., 2018; Isensee et al., 2018; Zhang et al., 2021b; Hatamizadeh et al., 2022; Mai et al., 2022). While the latter uses modality-specific encoders to extract individual features that are later combined in the middle layers of the network and share the same decoder (Xing et al., 2022; Zhang et al., 2021a; Ding et al., 2021; Zhou et al., 2020a, 2022; Dolz et al., 2018).

However, redundant information is present among medical images of different modalities, as evidenced by many works on the cross-modal generation that aim to increase the number of training samples or reduce medical costs by generating high-cost modalities based on low-cost modalities (Van Tulder and de Bruijne, 2015; Ben-Cohen et al., 2019; Zhou et al., 2020b; Bouman et al., 2023). In MML, taking into account that fixed representations can only encapsulate a limited amount of information, redundant information can cause joint learning algorithms to misjudge the importance of different modalities (Li et al., 2020), disregard specific modal information (Wang et al., 2022), generate additional cognitive load (Mayer and Moreno, 2003; Cao et al., 2009; Knoop-van Campen et al., 2019), and ultimately reduce prediction accuracy and result in overfitting (Lin et al., 2021; Mai et al., 2022).

Redundant information in MML can be categorized into *intra-modality* and *inter-modality* redundancy. The former refers to redundancy within a single modality, and the latter refers to consistent information across different modalities. Existing MML techniques predominantly focus on addressing *intra-modality* redundancy (Lin et al., 2021; Wang et al., 2022). In the case of *inter-modality* redundancy, two separate operations, namely *addition* and *subtraction*, can be utilized to minimize redundant information in joint representations, as depicted in Figure 1. A considerable number of mid-term fusion strategies (Xing et al., 2022; Zhang et al., 2021a; Ding et al., 2021; Zhou et al., 2020a, 2022; Dolz et al., 2018) implicitly reduce *inter-modality* redundancy during joint learning by integrating modalities and applying end-to-end learning principles. An alternative method (Mai et al., 2022) employs the information bottleneck to decrease redundancy in joint representations associated with the *subtraction* operation. Conversely, the *addition* operation, which amalgamates complementary information from multiple modalities, is intuitively superior efficacy in eradicating *inter-modality* redundant information in comparison to the *subtraction* operation.

Furthermore, experienced radiologists often analyze multimodal data in clinical practice by designating a primary modality and several auxiliary modalities for pathological diagnosis. This approach is exemplified in the BraTS challenge (Menze et al., 2014) (Figure 2). In this challenge, the segmentation target was the different areas of glioblastoma, which is the most common type of brain malignancy. Glioblastoma is characterized by its resistance to treatment and poor prognosis, making it a critical focus for advancements in medical imaging and treatment strategies(Li et al., 2019). Human annotators primarily employ the T2 modality[2] for segmenting the edema region while using the FLAIR modality to verify the presence of edema and other fluid-filled structures. Subsequently, the tumor core (TC) is identified through the combined use of T1CE and T1 modalities. The expertise of these radiologists suggests that specific mapping relationships exist between modalities and target areas. Certain modalities facilitate the identification of particular area boundaries, while others serve as supplementary aids. Gleaning insights from this expert knowledge and incorporating it as an inductive bias can potentially reduce the complexity of learning

---

2. T1, T1CE, T2, and FLAIR represent four modalities generated by MRI imaging technology.

relationships between modalities and corresponding regions. A similar example of adding priors to reduce learning difficulty is DetexNet (Liu et al., 2020), which simplifies low-level representation patterns by embedding expert knowledge.

Inspired by these observations, we propose the **Complementary Information Mutual Assistance Learning (CIML)** framework. The primary objective of the CIML framework is to effectively eliminate *inter-modal* redundant information in multi-modal segmentation tasks. To leverage doctors' prior knowledge regarding the correspondence between modalities and regions, the learning multimodal segmentation task is decomposed into multiple single-modal segmentation subtasks, as illustrated in Figure 3. Within the CIML framework, each modality serves a dual purpose: as the main modality, the corresponding segmentor encodes its information, integrates messages transmitted by other modalities, and extracts non-redundant information to complete the single-modal subtask; as an auxiliary modality, the corresponding segmentor transmits auxiliary information to other modalities. This approach minimizes the mutual influence of modal information and ensures sufficient feature extraction for the target area in each subtask. When multiple regions require segmentation, such as in `BraTS2020`, expert prior knowledge[3] can be employed to match modalities to regions. In tasks involving only one region to segment, such as `autoPET`, each modal segment is matched to the corresponding region individually. The primary mode is the matching mode of the target (sub)region, while the remaining modes serve as auxiliary modes. The final segmentation is obtained by combining or averaging all unimodal segmentations.

After task decomposition, since the auxiliary mode typically contains less additionally useful information (non-redundant information), we filter redundant information to extract non-redundant information representation from messages transmitted in the auxiliary mode. We first adopt an information theory perspective to transform the problem into an equivalent complementary information learning problem. Inspired by the variational information bottleneck (Alemi et al., 2017), we model this problem as a mutual information bi-objective optimization problem. Variational inference is then employed to make optimization problems more tractable, including cross-entropy and Kullback-Leibler (KL) divergence minimization, which can be efficiently solved through automatic differentiation. Finally, we introduce cross-modal spatial attention as a parameterized backbone to achieve practical implementation.

Overall, CIML adopts two mechanisms, namely *task decomposition* and *redundancy filtering*, to minimize the *inter-modal* redundant information that is relied upon by the algorithm in the segmentation process. Task decomposition physically minimizes the inter-dependence of information between modalities (through inductive bias). At the same time, redundancy filtering extracts as little information as possible from other modalities using the *addition* operation at the algorithm level. To validate the effectiveness of CIML, we first perform a visual examination of a human-designed image segmentation task to confirm its redundancy-free nature. Subsequently, we evaluate our framework on standard multi-modal medical image segmentation benchmarks, including `BarTS2020`, `autoPET`, and `MICCAI HECKTOR 2022`. Experimental results demonstrate that CIML significantly outperforms state-of-the-art algorithms in terms of validation accuracy and segmentation quality.

---

3. To study the robustness of CIML to different task decompositions, we conducted experiments on a medical image segmentation task where expert prior knowledge is not available, using random matching of modalities and regions to be segmented. Although this scenario is rare, experimental results show that CIML is moderately sensitive to different task decompositions.

Moreover, the incorporation of task decomposition and redundancy filtering allows us to utilize neural network visualization techniques, such as Grad-CAM (Selvaraju et al., 2017), to gain insights into the contribution of each modality to the segmentation of different regions. By visualizing the relationships and knowledge sharing among modalities, we enhance the credibility and interpretability of multimodal medical image segmentation algorithms, ultimately improving their effectiveness in clinical diagnosis and treatment. Main contributions can be summarized as follows:

1) We introduce the **Complementary Information Mutual Learning (CIML)** framework, which aims to enhance the information fusion efficiency in multimodal learning. CIML presents a pioneering approach by algorithmic modeling and mitigating the negative impact of *inter-modal* redundant information that arises in the joint learning used by state-of-the-art techniques;

2) CIML adopts a unique perspective of *addition* to eliminate *inter-modal* redundant information through inductive bias-driven **task decomposition** and message passing-based **redundancy filtering**, thus effectively decreasing the difficulty of constructing knowledge relationships among modalities in multimodel learning;

3) We establish an equivalent transformation from the redundant filtering problem to the complementary information learning problem based on the variational information bottleneck and solve it efficiently with variational inference and cross-modal spatial attention;

4) Message passing-based redundancy filtering allows for applying neural network visualization techniques, such as Grad-CAM (Selvaraju et al., 2017), for visualizing the knowledge relationship among different modalities, which reflects the interpretability.

The following is the roadmap of this paper. Section 2 provides the related works and preliminaries. Section 3 describes the proposed framework. Experiments and numerical results are presented in Section 4, and we conclude this paper in Section 6.

2. Related Work and Preliminaries

In this section, we present a comprehensive literature review on the state-of-the-art multimodal fusion strategies, mutual information, and information bottleneck techniques. In addition, we apply the class activation map methodology to visualize complementary information and thus provide an overview of this technique.

2.1 Multimodal Fusion and Redundancy Reducing

Multimodal machine learning has a broad range of applications, including but not limited to audio-visual speech recognition (Yuhas et al., 1989), image captioning (Xu et al., 2015), visual question answering (Wu et al., 2017), besides medical image analysis.

Multimodal learning involves the challenge of combining information from two or more modalities to perform accurate predictions (Baltrušaitis et al., 2018; Zhao et al., 2024). To effectively extract relevant information from multiple sources, various techniques must be employed to capture and integrate an appropriate set of informative features from multiple modalities. Early fusion and intermediate fusion schemes are the most commonly used methods for this purpose. Early fusion approaches (Oktay et al., 2018; Isensee et al., 2018; Zhang et al., 2021b; Hatamizadeh et al., 2022; Li et al., 2023) adopt a single stream

fusion strategy, where multimodal images fuse before input into a neural network. However, these methods can hardly explore the inter-modality connections. Intermediate fusion approaches(Xing et al., 2022; Zhang et al., 2021a; Ding et al., 2021; Zhou et al., 2020a, 2022; Dolz et al., 2018; Yao et al., 2024) follow a multi-stream fusion strategy, where features are fused in the middle layers of the network and share the same decoder. Among these multi-stream methods, attention mechanisms are often utilized to emphasize contributions from different modalities. Methods such as NestedFormer (Xing et al., 2022), ModalityNet (Zhang et al., 2021a), Tri-attentionNet (Zhou et al., 2022), and One-shotMIL (Zhou et al., 2020a) leverage attention mechanisms to achieve this. The Tri-attentionNet algorithm (Zhou et al., 2022) additionally models the relationship between modalities' features, which helps to improve segmentation accuracy. However, they do not fully utilize the relationship between tumor regions and modalities. To address this limitation, the RFNet framework (Ding et al., 2021) was proposed, which employs a region-aware fusion scheme. This approach considers the different contributions of various modalities to each region, as different modalities have distinct presentations and sensitivities to different tumor regions. Besides early fusion and intermediate fusion approaches, the PolicyFuser (Huang et al., 2023) retains one independent decision for each sensor and fusion decision. DrFuse (Yao et al., 2024) disentangles shared and unique features, then applies a disease-wise attention layer for each modality to make the final prediction.

Addressing missing modalities in multimodal medical image segmentation has emerged as a critical research focus. A common strategy involves synthesizing the missing modalities using generative models. For example, Orbes-Arteaga et al. (2018) demonstrated the synthesis of FLAIR images from T1 modality for segmentation tasks. Similarly, the heteromodal variational encoder-decoder framework proposed by Dorent et al. (2019) performs joint modality completion and segmentation, generating the required modality. Another approach focuses on learning modality-invariant feature spaces. The method by Havaei et al. (2016) achieves segmentation by learning features that are consistent across different modalities. ShaSpec (Wang et al., 2023a) further enhances this by leveraging all available modalities during training, learning both shared and specific features. Knowledge distillation and feature transfer techniques also address missing modalities effectively. The Modality-Aware Mutual Learning framework (Zhang et al., 2021a) involves modality-specific models learning collaboratively to distill knowledge. ProtoKD (Wang et al., 2023b) transfers pixel-wise knowledge from multi-modality to single-modality data, enabling robust feature representation and effective inference with a single modality. Latent space information imputation provides an alternative by estimating task-related information of missing modalities. M^{3}Care (Zhang et al., 2022) utilizes auxiliary information from similar patient neighbors, guided by a modality-adaptive similarity metric, to estimate missing data in the latent space, thus facilitating clinical tasks. Finally, self-attention-based modality fusion offers another solution. SFusion (Liu et al., 2023) introduces a self-attention-based fusion block that automatically learns to fuse available modalities without synthesizing missing ones. By projecting features into a self-attention module and generating latent multimodal correlations, SFusion constructs a shared representation for downstream models.

These varied approaches underscore the innovative strategies developed to handle missing modalities in multimodal medical image segmentation, enhancing the robustness and accuracy of segmentation models. In our work, we aim to eliminate redundancy to extract more

effective information, which contrasts with most methods addressing missing modalities. These methods typically utilize redundant information to complete other modalities, enabling the use of networks designed for full modalities to accomplish segmentation tasks. The main strategy of these works is contrary to ours. Moreover, the M³Care (Zhang et al., 2022) method is worth considering for integration with our approach. By leveraging similar patient neighbors, M³Care can extract the necessary complementary information to aid in segmentation. In future work, we will try to learn from M³Care to enable our method to cope with the situation of missing modalities.

Further, the current methods for fusing multimodal features do not consider *inter-modal* redundancy, which may lead to misjudging the importance of modalities, ignoring specific modal information, and increasing cognitive load. To address this limitation, two methods have been proposed in the context of multi-view, which is similar to multimodal data. CoUFC (Zhao et al., 2020) couples the correlated feature matrix and the uncorrelated ones together to reconstruct data matrices. Although CoUFC utilizes an implicit way of eliminating redundancy by focusing on correlated features and uncorrelated features, its solution is not applicable to high-dimensional, high-resolution medical images. Another work (Tosh et al., 2021) introduces a contrastive learning method, which learns transformation functions from one view to the other in an unsupervised fashion and then learns a linear predictor for downstream tasks. However, this work focuses on theoretical analysis, lacks validation on complex high-dimensional data, and does not eliminate intra-modal redundancy due to its unsupervised fashion.

There are two main differences between CIML and existing approaches. Firstly, our CIML algorithm decomposes the original task, thereby facilitating the establishment of an association between modal and target regions. Secondly, we employ redundancy filtering to extract complementary information, thereby eliminating redundancy and maximizing information gain from auxiliary modalities.

2.2 Mutual Information and Information Bottleneck

The application of information-theoretic objectives to deep neural networks was first introduced in Tishby and Zaslavsky (2015), although it was deemed infeasible at the time. However, variational inference provides a natural way to approximate the problem. To bridge the gap between traditional information-theoretic principles and deep learning, the variational information bottleneck (VIB) framework was proposed in Alemi et al. (2017). This framework approximates the information bottleneck (IB) constraints, enabling the application of information-theoretic objectives to deep neural networks.

Several works (Federici et al., 2020; Wu and Goodman, 2018; Zhu et al., 2020; Lee and van der Schaar, 2021; Mai et al., 2022; Wang et al., 2019) have been proposed to adopt the IB for multi-view or MML, which is the most relevant to our work. IB variants such as those proposed in Federici et al. (2020); Wu and Goodman (2018); Zhu et al. (2020); Lee and van der Schaar (2021) extend the VIB framework for multi-view learning. These methods obtain a joint representation via Product-of-Expert (PoE) (Hinton, 2002). Another work (Wang et al., 2019) proposes a deep multi-view IB theory, which aims to maximize the mutual information between the labels and the learned joint representation while simultaneously minimizing the mutual information between the learned representation of each view and the

original representation. In addition, a recent study (Mai et al., 2022) introduced a multimodal IB approach, which aimed to learn a multimodal representation that is devoid of redundancy and can filter out extraneous information in unimodal representations. Instead of applying PoE, this work develops three different IB variants to study multimodal representation. CIML differs from the above multi-view IB methods in the following ways: 1). We take into account the varying importance of different modalities. Drawing on expert knowledge that specific modalities contain a greater amount of relevant information than others, CIML decomposes the task and designates some modalities as primary and others as auxiliary; 2). We assume that the primary modality contains the majority of information about the target region. To maximize information gain and minimize redundancy for segmentation, our method constrains the representation from the auxiliary modalities to contain only complementary information.

2.3   Class Activation Map

A widely-used method for determining the most influential pixels or voxels, specifically those with intensity changes that significantly affect the prediction score, involves the generation of a class activation map (CAM) (Zhou et al., 2016; Selvaraju et al., 2017). These maps highlight the regions in the input data that contribute the most to the model's output, thereby providing insights into the decision-making process of the model. CAM assigns weights to feature maps in a specific convolutional layer and can be easily integrated into a pre-trained deep model without introducing additional parameters. Several variations have been proposed that build upon CAM to more accurately highlight important regions in the image, such as Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). Grad-CAM uses the gradient signal of the activations in a convolutional layer and has been successfully applied to image classification. An extension of Grad-CAM (Vinogradova et al., 2020) produces heatmaps showing the relevance of individual pixels or voxels for semantic segmentation.

In this work, we apply Grad-CAM to visualize voxels that provide complementary information on auxiliary modalities. By doing so, we aim to identify and highlight the most informative and relevant regions in the auxiliary modality for accurate target prediction.

3. Methodology

In this section, we introduce the Complementary Information Mutual Learning (CIML) framework for medical image segmentation, which aims to efficiently segment through eliminating *inter-modality* redundancy. The framework incorporates two primary mechanisms: *task decomposition* (Section 3.1) and *redundancy filtering* (Section 3.2). The *task decomposition* mechanism seeks to reduce the interdependence of information between modalities by drawing on expert prior knowledge as an inductive bias. On the other hand, the redundancy filtering mechanism reduces the amount of redundant information extracted from other modalities through the variational information bottleneck and variational inference. We also introduce the *cross-modality information gate module* that utilizes cross-modal spatial attention to implement redundancy filtering practically.

We assume that our dataset comprises independent and identically distributed (i.i.d.) samples $\{X^i \in \mathbb{R}^{T \times W}\}_{i=1}^{N}$ drawn from a medical image data distribution. In this context, i

represents the index of the image, while T and W represent the number of modalities and the number of voxels, respectively. To distinguish between different modalities, we employ subscripts, such as $\{X_m^i\}_{m \in \{1,2,\cdots,T\}}$. Our objective is to segment the image into u distinct regions by classifying every voxel into one of u classes. We define $Y^i \in \{1, 2, \cdots, u\}^W$ as the segmentation mask for the i-th sample. Since the multimodal images are spatially aligned, all modalities within a single sample share the same mask, which can be expressed as $Y_m^i = Y^i$ for all $m \in \{1, 2, \cdots, T\}$.

3.1 Inductive Bias-driven Task Decomposition

CIML applies a unique perspective of *addition* to eliminate *inter-modal* redundant information. The first step of *addition* is *task decomposition*, which is driven by the inductive bias extracted from expert prior knowledge. As shown in Fig. 3, *task decomposition* involves decomposing the task into several subtasks. For sub-task τ_ω, the modality X_{γ_ω} is assigned as the primary modality, which contains significant information for the target (sub-)regions segmentation. In some cases, multiple modalities are combined as primary modalities for a sub-task, depending on the task. Furthermore, the remaining modalities, which serve as primary modalities for other sub-tasks, are treated as auxiliary modalities that provide complementary information to assist with the sub-task τ_ω. This unique perspective of *addition* allows us to exploit the complementary information from multiple modalities effectively, reducing redundancy and improving the accuracy and efficiency of the segmentation algorithm.
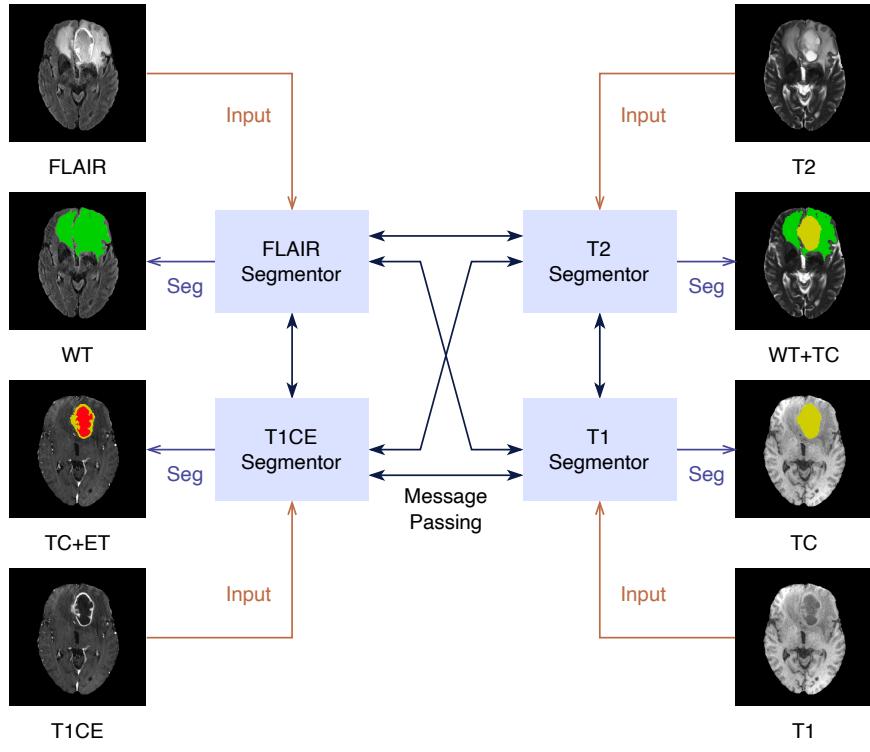
Additionally, we utilize a message-passing mechanism between sub-tasks to transport efficient information from auxiliary modalities. CIML utilizes a distinct sub-model for each sub-task, which is responsible for extracting uni-modal features, performing message passing, obtaining complementary information, and predicting the target (sub-)regions. These sub-models are referred to as *segmentors* (Figure 4), with the segmentor for sub-task τ_ω denoted as

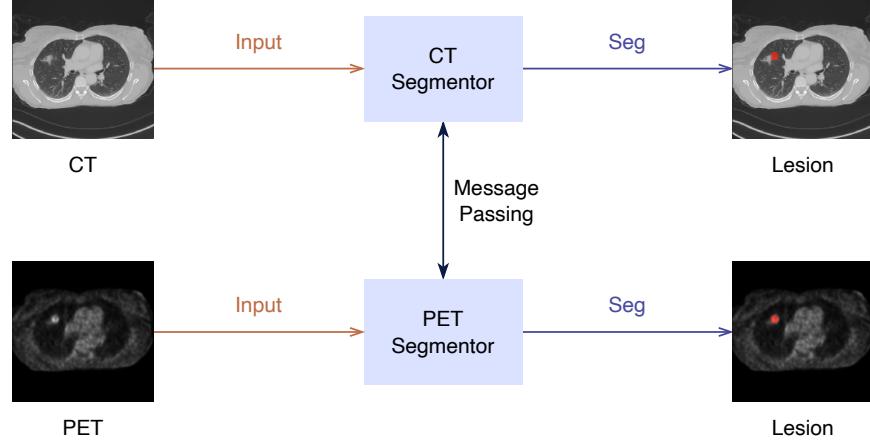$$f_\omega \left(X_{\gamma_\omega}, M_{1 \sim T/ \; \gamma_\omega} \mid \theta_\omega \right)$$

and parameterized by θ_ω. In this notation, $M_{1 \sim T/ \; \gamma_\omega}$ denotes messages from other sub-tasks.

A segmentor comprises three fundamental components: a message generator, a complementary information filter, and a predictor. The message generator encodes input images to produce embeddings and messages. Subsequently, the complementary information filter utilizes the embeddings from the message generator and messages from other modalities to extract complementary information, thereby enhancing the segmentor's performance. Lastly, the predictor leverages the embeddings generated by the message generator and the complementary information to generate the final predictions.

To describe *task decomposition* more clearly, we take BraTS2020 as an example. As illustrated in Fig.3, the original multi-target task is divided into four distinct sub-tasks. The sub-tasks involve utilizing FLAIR as the primary modality for segmenting the whole tumor (WT) region, employing T1 as the primary modality for segmenting both the tumor core (TC) and the enhanced tumor (ET), leveraging T2 as the primary modality for segmenting the WT and TC regions and adopting T1CE as the primary modality for segmenting the TC and ET regions. This task decomposition is based on expert prior knowledge, which suggests that the selected primary modalities contain the most informative features for accurately segmenting their respective target regions. In addition, to test the performance of different

(a) CIML for `BraTS2020` challenge.

(b) CIML for `autoPET` challenge.

Figure 3: Illustration of complementary information mutual learning (CIML) framework for `BraTS2020` challenge and `autoPET` challenge. The input to each segmentor consists of multimodal images that are specific to each modality. After processing, the segmentors send a portion of the embeddings as messages to other segmentors to assist with other sub-tasks and accept messages from other segmentors to extract efficient information. The dark blue lines with bi-directional arrows in the figures represent the message passing. Finally, the segmentors complete their sub-tasks. The `MICCAI HECKTOR 2022` challenge also applies a similar framework to the `autoPET` challenge.
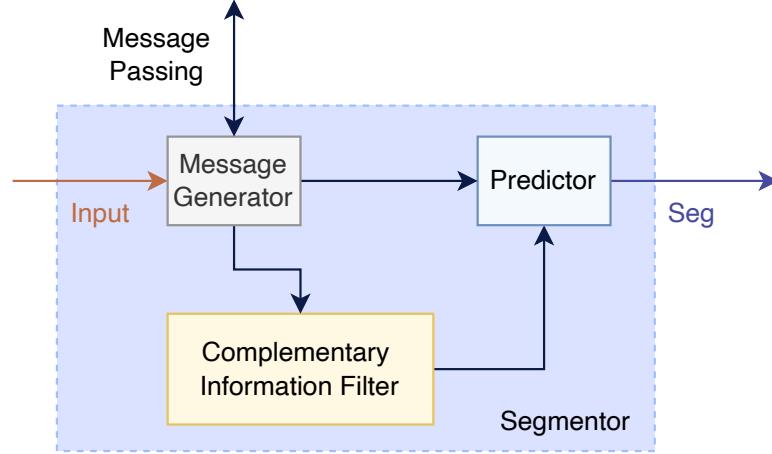
Figure 4: Segmentor contains three parts: message generator, complementary information filter, and predictor.
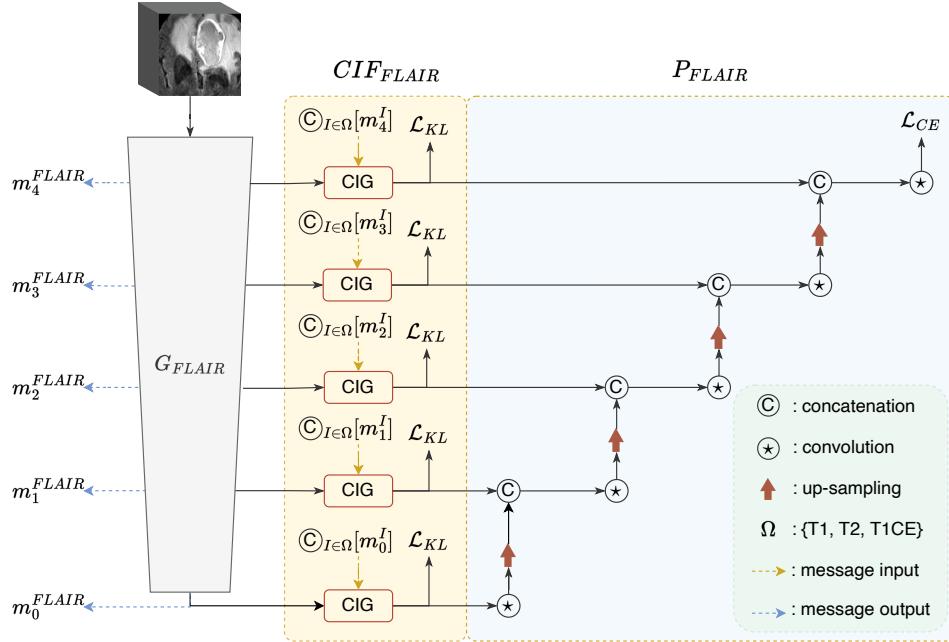

Figure 5: Schematic of the network architecture of the segmentor. Generator G_{FLAIR} is employed to extract features from FLAIR images individually. Complementary Information Filter (CIF) Module is used to extract complementary information from messages, and predictor P_{FLAIR} is utilized to generate the final segmentation.

decompositions, we designed ablation experiments to compare several different decomposition methods, as detailed in Section 4.

The segmentor architecture is based on the nnUNet (Isensee et al., 2018) and utilizes an intermediate fusion scheme. It consists of an encoder and a decoder, which are connected

through skip connections. A schematic representation of the segmentor's overall structure is provided in Fig.4, while the structure corresponding to sub-task τ_{FLAIR} is depicted in Fig.5. For the sub-task τ_{FLAIR}, the message generator, G_{FLAIR}, acts as the encoder and is composed of four stages. It takes the FLAIR images as input, which are first cropped into 3D patches and generate embedded images. At each stage, G_{FLAIR} produces embeddings that are skip-connected to the decoder in the original U-Net architecture. These embeddings serve as messages denoted by $\{M^{\omega}_{FLAIR}\}_{\omega \in \{1,2,3,4\}}$ for assisting other sub-tasks. The decoder of the segmentor for sub-task τ_{FLAIR} is composed of the complementary information filter (CIF_{FLAIR}) and predictor (P_{FLAIR}). CIF_{FLAIR} is designed to utilize the embeddings from the message generator (G_{FLAIR}) and messages from other segmentors to filter out complementary information. It also contains four stages, and in each stage, CIF_{FLAIR} incorporates a Cross-modality Information Gated (CIG) module based on cross-modal spatial attention, which will be explained in detail in the following subsection. The output of CIF_{FLAIR} combines the embeddings from the encoder and the complementary representation. Finally, P_{FLAIR} predicts the results of the segmentor based on the output of the complementary information filter.

3.2 Message Passing-based Redundancy Filtering

In the second phase of the *addition* process, message passing-based redundancy filtering is employed to eradicate *inter-modality* redundancy, thereby extracting supplementary information. Drawing inspiration from the variational information bottleneck, we reformulate the problem as a bi-objective mutual information optimization problem. Subsequently, we leverage variational inference to make the optimization problem more tractable by minimizing cross-entropy and KL divergence, and we efficiently solve it using automatic differentiation. Finally, we employ cross-modal spatial attention as a parameterization backbone to obtain a practical implementation.

To facilitate discussion, we concern a scenario where two sub-tasks, specifically $\tau_1 : X_1 \rightarrow Y_1$ and $\tau_2 : X_2 \rightarrow Y_2$, are present. Our primary focus is on sub-task τ_1, and the same principles can be extended to more than two sub-tasks. In this context, X_1 and X_2 signify the first and second modalities, while Y_1 and Y_2 denote the corresponding ground truth for each sub-task.

X_1 represents the primary modality encompassing the majority of information pertaining to the target region Y_1. In accordance with standard supervised learning literature, we predict Y_1 directly by minimizing the supervised learning loss:

$$\mathcal{L}_{SL} = -\mathbb{E}_{x_1^i, y_1^i \sim \{X_1, Y_1\}} \log f_{X_1}(y_1^i \mid x_1^i), \tag{1}$$

where f_{X_1} denotes the parameterized segmentation function. Besides, we aim to generate a representation K_2 derived from primary and auxiliary modalities that encapsulate complementary information to aid in predicting the target Y_1.

This problem can be modeled within a Bayesian graph (refer to Figure 6) following three Markov chains: $X_1 \leftarrow X \rightarrow X_2$, $X_1 \rightarrow Y_1 \leftarrow X_2$, and $X_1 \rightarrow K_2 \leftarrow X_2$. Here, X represents the patient as a hidden variable, and K_2 is the representation derived from X_1 and X_2.
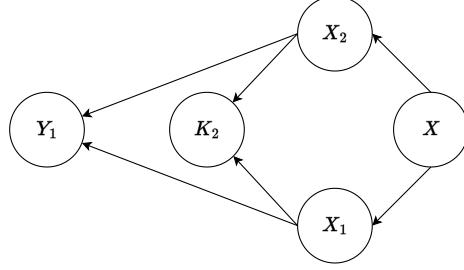
Figure 6: A Bayes graph representing the relationship of modalities X_1, X_2, target Y_1, and representation K_2.
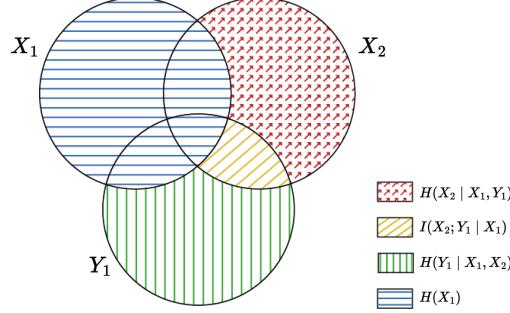

Figure 7: Venn diagram of information-theoretic measures for three variables X_1, X_2, and Y_1, represented by the upper left, upper right, and lower circles, respectively.

Figure 7 illustrates the interdependence among variables X_1, X_2, and the target variable Y_1 using a Venn diagram to demonstrate their overlapping relationships (see B.1 for some basic knowledge about Mutual Information and Venn Diagrams). Entropy (H) and mutual information (I) are essential concepts in information theory, as they measure the uncertainty in a set of outcomes and the reduction in uncertainty about one variable due to the knowledge of another variable, respectively (Shannon, 2001). Further, representation K_2 is generated following the Markov chain $X_1 \rightarrow K_2 \leftarrow X_2$, so the Venn diagram of K_2 is within the union of Venn diagram of X_1 and X_2. The area shown in Figure 7 with yellow stripes represents the complementary information not contained in X_1 and contributes to the identification of Y_1.

Employing Markov chains as a foundation, the mutual information $\mathcal{I}_1(X_1, X_2; K_2)$ can be partitioned into three distinct components through the application of the chain rule of mutual information (more details see Appendix C.1):

$$
\begin{aligned}
&\mathcal{I}_1(X_1, X_2; K_2) \\
&= \underbrace{\mathcal{I}_2(K_2; Y_1 \mid X_1)}_{\text{complementary predictive information}} + \underbrace{\mathcal{I}_3(K_2; X_1)}_{\text{duplicated information}} \\
&+ \underbrace{\mathcal{I}_4(K_2; X_2 \mid X_1, Y_1)}_{\text{unique but irrelevant information}},
\end{aligned}
\tag{2}
$$

15

where $\mathcal{I}_2(K_2; Y_1 \mid X_1)$ represents the information in K_2 that is not involved by modality X_1 and is predictive of Y_1. While $\mathcal{I}_3(K_2; X_1)$ indicates the duplicated information already involved in modality X_1, and $\mathcal{I}_4(K_2; X_2 \mid X_1, Y_1)$ indicates unique information but irrelevant information. We regard $\mathcal{I}_3$ and $\mathcal{I}_4$ as *inter-modality* redundancy.

Based on these observations, we formulate two objectives to generate K_2, i.e.,

$$\begin{cases} \min_{K_2} \mathcal{I}_1(X_1, X_2; K_2), \\ \max_{K_2} \mathcal{I}_2(K_2; Y_1 \mid X_1). \end{cases} \tag{3}$$

The first objective seeks to maximize the mutual information between K_2 and the target Y_1, given the modality information of modality X_1. This constraint ensures K_2 contains the information depicted in the Venn diagram with yellow stripes. To guarantee that K_2 encompasses solely essential information and minimizes redundancy, the mutual information between K_2 and modalities X_1, X_2 is minimized. This dual objective optimization constrains K_2 to be a representation of complementary information containing only indispensable information. Considering the above Bayesian network, the joint probability can be expressed as (see detailed derivation in Appendix B.2):

$$p(X_1, X_2, Y_1, K_2) = p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2, Y_1). \tag{4}$$

Furthermore, variational inference can be employed to render the optimization problem unconstrained. The first objective has an upper bound (see detailed derivation in Appendix C.2):

$$\begin{aligned} \mathcal{I}_1 &(X_1, X_2; K_2) \\ &= \int p\left(x_1, x_2, \kappa_2\right) \cdot \log \left(p\left(\kappa_2 \mid x_1, x_2\right) / p(\kappa_2)\right) dx_1 d\kappa_2 dx_2 \\ &\leq \int p(x_1, x_2, \kappa_2) \cdot \log \left(p\left(\kappa_2 \mid x_1, x_2\right) / r(\kappa_2)\right) dx_1 d\kappa_2 dx_2 \\ &\approx \frac{1}{N} \sum_i^N \int p\left(\kappa_2 \mid x_1^i, x_2^i\right) \cdot \log \left(p\left(\kappa_2 \mid x_1^i, x_2^i\right) / r(\kappa_2)\right) d\kappa_2, \end{aligned} \tag{5}$$

where $r(\kappa_2)$ is a standard normalization distribution. In practice, we can use neural networks $\mu_\theta(x_1, x_2), \sigma_\theta(x_1, x_2)$ to approximate $p\left(\kappa_2 \mid x_1, x_2\right)$ by $\mathcal{N}\left(\mu_\theta(x_1, x_2), \sigma_\theta(x_1, x_2)\right)$. The second mutual information maximization objective possesses a lower bound (see detailed derivation in Appendix C.3):

$$\begin{aligned} \mathcal{I}_2 &(K_2; Y_1 \mid X_1) \\ &= \int dx_1 d\kappa_2 dy_1 dx_2 p(x_1, x_2, \kappa_2, y_1) \log \frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)} \\ &\geq \int dx_1 d\kappa_2 dy_1 dx_2 p(x_1, x_2, \kappa_2, y_1) \log \frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)} \\ &\approx \frac{1}{N} \sum_i^N \int d\kappa_2 p(\kappa_2 \mid x_1^i, x_2^i) \log q(y_1^i \mid \kappa_2, x_1^i) + H, \end{aligned} \tag{6}$$

16

where $q(y_1 \mid \kappa_2, x_1)$ serves as a variational approximation to $p(y_1 \mid \kappa_2, x_1)$. Notice that H is independent of the optimization procedure and can be ignored. Combining both of these bounds, we have that,

$$
\mathcal{L}_{com} \approx \frac{1}{N} \sum_{i=1}^{N} \int \left[-p\left(\kappa_2 \mid x_1^i, x_2^i\right) \cdot \log q\left(y_1^i \mid \kappa_2, x_1^i\right) \right.
$$
$$
\left. + \beta\, p\left(\kappa_2 \mid x_1^i, x_2^i\right) \cdot \log\left(p(\kappa_2 \mid x_1^i, x_2^i)/r(\kappa_2)\right) \right] d\kappa_2. \tag{7}
$$

Furthermore, we can formulate the loss in this way,

$$
\mathcal{L}_{com} = \mathcal{L}_{CE} + \beta \mathcal{L}_{KL}, \tag{8}
$$

where $\beta \geq 0$ controls the tradeoff between two objectives. The total loss contains cross entropy and Kullback–Leibler divergence, and the former is

$$
\mathcal{L}_{CE} = \frac{1}{N} \sum_{i=1}^{N} \int \left[-p(\kappa_2 \mid x_1^i, x_2^i) \log q(y_1^i \mid \kappa_2, x_1^i) \right] d\kappa_2
$$
$$
= \mathbb{E}_{\substack{\epsilon \sim \mathcal{N}(0,1) \\ x_1^i, x_2^i, y_1^i \sim \{X_1, X_2, Y_1\}}} \quad - \log q\left(y_1^i \mid x_1^i, f_\theta(x_1^i, x_2^i, \epsilon)\right), \tag{9}
$$

where κ_2 is sampled with $f_\theta(x_1^i, x_2^i, \epsilon)$ which is a deterministic function of x_1^i, x_2^i and the gaussian random variable ϵ which is sampled from a normal gaussian distribution. On the other hand, $\mathcal{L}_{KL}$ is shown as follow

$$
\mathcal{L}_{KL} = \frac{1}{N} \sum_{i=1}^{N} \int \left[p\left(\kappa_2 \mid x_1^i, x_2^i\right) \cdot \log\left(p\left(\kappa_2 \mid x_1^i, x_2^i\right)/r(\kappa_2)\right) \right] d\kappa_2
$$
$$
= \mathrm{KL}\left[p\left(K_2 \mid X_1, X_2\right) \| r(K_2) \right]. \tag{10}
$$

Since both $q(y_1^i \mid \kappa_2, x_1^i)$ and $f_{X_1}(x_1^i)$ involve X_1 as input and share the same prediction target Y_1, the cross-entropy loss $\mathcal{L}_{CE}$ and the supervised loss $\mathcal{L}_{SL}$ can be simplified by combining them and retaining only the cross-entropy loss $\mathcal{L}_{CE}$.

The resulting loss function consists of two components: the cross-entropy term measures the discrepancy between our predictions and the targets, while the KL term constrains the representations from CIG modules to represent complementary information. This approach enables the efficient extraction of complementary information representations from messages through automatic differentiation. The function $q\left(y_1^i \mid x_1^i, f(x_1^i, x_2^i, \epsilon)\right)$ can be utilized to predict the target Y_1.

Moreover, in order to improve the extraction of complementary information in auxiliary modalities, we propose a *cross-modality information gate* (CIG) that merges our formulated loss function. The CIG module utilizes cross-modal spatial attention as a parameterization backbone to achieve a practical implementation.

Figure 5 shows that each stage of the complementary information filter in the segmentor contains a CIG module. This CIG module is based on a cross-modal spatial attention mechanism and our proposed loss in Equation 8 to extract complementary information. Figure 8 shows the architecture of the CIG module. It inputs the output features of encoders
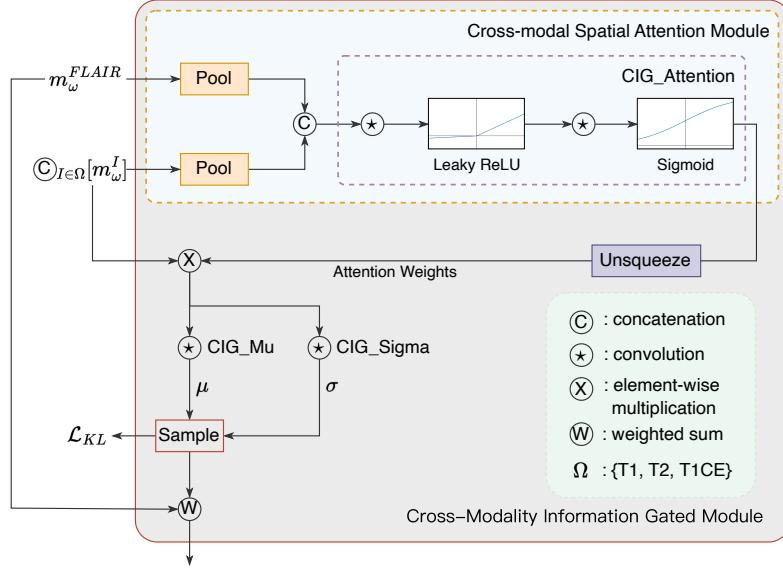
Figure 8: Visualization of the Cross-Modality Information Gated Module: utilizing cross-modal spatial attention to identify key voxels, filtering complementary information via attention weights, and incorporating residual mechanism for combining local information.

and the messages from other segmentors. The spatial attention mechanism has proven effective in achieving high performance with limited parameters and has been utilized in various computer vision applications (Woo et al., 2018; Zhou et al., 2022; Hinton et al., 2015). We leverage a cross-modal spatial attention module to extract complementary information that is only contained in messages from other segmentors.

The CIG module utilizes average pooling to reduce the number of network parameters while preserving location information. The pooled features are then concatenated and squeezed before passing through CIG_Attention module, which contains two convolutional layers with Leaky ReLU and sigmoid activation functions to obtain the cross-modality attention weights. These weights highlight the critical voxels and are unsqueezed and multiplied element-wise with the messages from other sub-tasks.

To generate the complementary information features, two convolutional layers, CIG_Mu and CIG_sigma, are utilized to obtain the mean μ and standard deviation σ. These parameters are then used in the reparameterization trick, which is constrained by KL terms. We employ a weighted sum operation to integrate dominant and complementary information features. Specifically, we first align the channel dimensions of the auxiliary modality features with those of the primary modality by applying convolutional (weighted) adjustments. Subsequently, we combine these features through summation to complete the information.

With this module, each segmentor in the proposed CIML framework can effectively extract complementary information from auxiliary modalities, allowing for the improvement of multimodal medical image segmentation.

18

4. Experiments and Results

In this section, we investigate two questions to determine the feasibility and efficiency of our approach:

Q1): Can the message passing-based redundancy filtering effectively extract non-redundant information from messages transmitted by auxiliary modalities?

Q2): Whether removing *inter-modality* redundancy can improve the quality of medical image segmentation?

To solve these issues, we evaluate the proposed approach on four tasks, namely the `ShapeComposition`, `BraTS2020`, `autoPET` and the `MICCAI HECKTOR 2022`. We address Q1) using the hand-crafted demonstrated task `ShapeComposition` and address Q2) using three standardized benchmarks `BraTS2020`, `autoPET` and `MICCAI HECKTOR 2022`. To further evaluate the effectiveness of our proposed CIML, we conduct ablation experiments by removing components from the proposed framework.

Unlike existing SOTA methods, task decomposition and redundant filtering enable us to use neural network visualizers, such as Grad-CAM (Selvaraju et al., 2017), to provide insight into the contribution of each modality to the segmentation of different regions. By visualizing the relationship of knowledge among modalities, the credibility of multimodal medical image segmentation algorithms is improved, enhancing their effectiveness in clinical diagnosis and treatment.

4.1 Public Dataset and Evaluation Metrics

4.1.1 Datasets

We evaluate the performance of the proposed CIML on both a demonstration task, named as `ShapeComposition`, and three publicly available datasets, namely `BraTS2020` (Menze et al., 2014), `autoPET` (Gatidis et al., 2022), and `MICCAI HECKTOR 2022` (Oreiller et al., 2022). `BraTS2020` is a brain tumor segmentation dataset consisting of four different modalities: Flair, T1CE, T1, and T2, while the `autoPET` and `MICCAI HECKTOR 2022` datasets contain positron emission tomography (PET) and computed tomography (CT) images, respectively.

The `BraTS2020` includes 369 subjects for training, with three distinct regions targeted for segmentation: the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET), in addition to the background. The `autoPET` challenge is composed of $1,014$ studies obtained from the University Hospital Tübingen and is publicly accessible on TCIA. The challenge aims to segment the lesion region. The `MICCAI HECKTOR 2022` dataset consists of 524 training cases collected from seven different centers, with the goal of segmenting images into two regions: background and lymph nodes (GTVn). For all datasets, we analyze the performance of various methods via five-fold cross-validation.

In addition, we manually decompose the segmentation task beforehand to investigate the effects of different assignments on the segmentation performance, which will be presented in the ablation study. For `BraTS2020`, we default set four sub-tasks, where FLAIR images are used to segment WT regions; T1 images are used to segment TC regions; T2 images are used to segment WT and TC regions; T1CE images are used to segment TC and ET regions. Since `autoPET` and `MICCAI HECKTOR 2022` contain only one target region, we set two segmentors that predict the same target region in these challenges. In the final results,

Table 1: Network configurations of our CIML. We report the Operators, Input Size, Output Size and Kernel Size. Stride can be easily inferred according to Input and Output size as we apply padding equals 1. Additionally, we use Batch Normalization in the BraTS2020 dataset and Instance Normalization in other experiments. In detail, we use P to denote patch size, C to denote the base number of filters, O to denote the number of channels of output, and K to denote the number of messages.

Architecture	Modules	Operators	Input Size	Output Size	Kernel Size
Encoder	Down1	Conv3D + Norm + LeakyReLU Dilated Conv3D + Norm + LeakyReLU	$P^3 \times 1$ $P^3 \times C$	$P^3 \times C$ $(P/2)^3 \times C$	3^3 3^3
	Down2	Conv3D + Norm + LeakyReLU Dilated Conv3D + Norm + LeakyReLU	$(P/2)^3 \times C$ $(P/2)^3 \times 2C$	$(P/2)^3 \times 2C$ $(P/4)^3 \times 2C$	3^3 3^3
	Down3	Conv3D + Norm + LeakyReLU Dilated Conv3D + Norm + LeakyReLU	$(P/4)^3 \times 2C$ $(P/4)^3 \times 4C$	$(P/4)^3 \times 4C$ $(P/8)^3 \times 4C$	3^3 3^3
	Down4	Conv3D + Norm + LeakyReLU Dilated Conv3D + Norm + LeakyReLU	$(P/8)^3 \times 4C$ $(P/8)^3 \times 8C$	$(P/8)^3 \times 8C$ $(P/16)^3 \times 8C$	3^3 3^3
Decoder	CIG_Attention	Conv3D + Norm + LeakyReLU Conv3D + Norm + Sigmoid	$(P/16)^3 \times (K+1)$ $(P/16)^3 \times 4(K+1)$	$(P/16)^3 \times 4(K+1)$ $(P/16)^3 \times K$	3^3 1^3
	[CIG_Mu]*k	Conv3D	$(P/16)^3 \times 8C$	$(P/16)^3 \times 8C$	1^3
	[CIG_Sigma]*k	Conv3D	$(P/16)^3 \times 8C$	$(P/16)^3 \times 8C$	1^3
	Up1	ConvTranspose3d + Norm + LeakyReLU	$(P/16)^3 \times 16C$	$(P/8)^3 \times 8C$	3^3
		Conv3D + Norm + LeakyReLU	$(P/8)^3 \times 16C$	$(P/8)^3 \times 8C$	3^3
	CIG_Attention	Conv3D + Norm + LeakyReLU Conv3D + Norm + Sigmoid	$(P/8)^3 \times (K+1)$ $(P/8)^3 \times 4(K+1)$	$(P/8)^3 \times 4(K+1)$ $(P/8)^3 \times K$	3^3 1^3
	[CIG_Mu]*k	Conv3D	$(P/8)^3 \times 4C$	$(P/8)^3 \times 4C$	1^3
	[CIG_Sigma]*k	Conv3D	$(P/8)^3 \times 4C$	$(P/8)^3 \times 4C$	1^3
	Up2	ConvTranspose3d + Norm + LeakyReLU Conv3D + Norm + LeakyReLU	$(P/8)^3 \times 8C$ $(P/4)^3 \times 8C$	$(P/4)^3 \times 4C$ $(P/4)^3 \times 4C$	3^3 3^3
	CIG_Attention	Conv3D + Norm + LeakyReLU Conv3D + Norm + Sigmoid	$(P/4)^3 \times (K+1)$ $(P/4)^3 \times 4(K+1)$	$(P/4)^3 \times 4(K+1)$ $(P/4)^3 \times K$	3^3 1^3
	[CIG_Mu]*k	Conv3D	$(P/4)^3 \times 2C$	$(P/4)^3 \times 2C$	1^3
	[CIG_Sigma]*k	Conv3D	$(P/4)^3 \times 2C$	$(P/4)^3 \times 2C$	1^3
	Up3	ConvTranspose3d + Norm + LeakyReLU Conv3D + Norm + LeakyReLU	$(P/4)^3 \times 4C$ $(P/2)^3 \times 4C$	$(P/2)^3 \times 2C$ $(P/2)^3 \times 2C$	3^3 3^3
	CIG_Attention	Conv3D + Norm + LeakyReLU Conv3D + Norm + Sigmoid	$(P/2)^3 \times (K+1)$ $(P/2)^3 \times 4(K+1)$	$(P/2)^3 \times 4(K+1)$ $(P/2)^3 \times K$	3^3 1^3
	[CIG_Mu]*k	Conv3D	$(P/2)^3 \times C$	$(P/2)^3 \times C$	1^3
	[CIG_Sigma]*k	Conv3D	$(P/2)^3 \times C$	$(P/2)^3 \times C$	1^3
	Up4	ConvTranspose3d + Norm + LeakyReLU Conv3D + Norm + LeakyReLU	$(P/2)^3 \times 2C$ $P^3 \times 2C$	$P^3 \times C$ $P^3 \times C$	3^3 3^3
	Output	Conv3D	$P^3 \times C$	$P^3 \times O$	3^3

we ensemble the results of each segmentor by averaging the results if the same target region is present in multiple segmentors. As `BraTS2020` is widely used in the literature, we mainly focus on this dataset in our experiments.

4.1.2 EVALUATION METRICS

The evaluation metrics in our experiments include the dice coefficient and 95% Hausdorff distance (HD):

- **Dice coefficient** (Dice, 1945): the dice coefficient measures the segmentation performance of CIML. Concretely, the dice coefficient from set X to set Y is defined as:

$$\mathrm{Dice}(\mathrm{X}, \mathrm{Y}) = \frac{2 \cdot \|\mathrm{X} \cap \mathrm{Y}\|_1}{\|\mathrm{X}\|_1 + \|\mathrm{Y}\|_1}. \tag{11}$$

It is worth highlighting that higher dice coefficients imply that the predictions are closer to the ground truth, which indicates more accurate segmentation.

- **HD95** (Henrikson, 1999): The maximum Hausdorff distance is the maximum distance of a set to the nearest point in the other set. More formally, The maximum Hausdorff distance from set X to set Y is defined as:

$$d_{\mathrm{H}}(X, Y) = \max \left\{ \sup_{x \in X} d(x, Y), \sup_{y \in Y} d(y, X) \right\}, \tag{12}$$

where sup represents the supremum, $d(a, B)$ is the shortest Euclidean distance between point a and set B.

4.2 Implementation Details

We run all experiments based on Python 3.8, PyTorch 1.12.1, and Ubuntu 20.04. All training procedures are performed on a single NVIDIA A100 GPU with 40GB memory. The initial learning rate is set to $1e-4$, and we employ a "poly" decay strategy as below:

$$\mathrm{lr} = \mathrm{initial\_lr} \times \left(1 - \frac{\mathrm{epoch\_id}}{\mathrm{max\_epoch}} \right)^{0.9}. \tag{13}$$

We apply Adam as the optimizer with weight decay set to $3e-5$ and betas set to (0.9, 0.999). We trained our models using a maximum number of epochs set to 500 (i.e., the max epoch in Equation 13) for the three public datasets and 1000 for the demonstration experiment. Each epoch consists of 100 iterations. To enhance the generalization of our models, we applied the default augmentation strategy in nnUNet (Isensee et al., 2018) for the three public datasets.

The configurations of the segmentor for the three public datasets are presented in Table 1. The network architecture for the demonstration experiment is similar but has fewer filters, which are described in more detail in Section 4.3. LeakyReLU activation with a negative slope of 0.01 is used in all experiments. Batch Normalization is used with a batch size of 2 in the `BraTS2020` dataset, while Instance Normalization is used in the other experiments. The

patch size is set to $64 \times 64 \times 64$ for the `BraTS2020` dataset and $128 \times 128 \times 128$ for other datasets, unless otherwise specified. The notation P refers to patch size, C refers to the base number of filters and K refers to the number of messages.

In our practical implementation, we utilize the sum of Dice loss (Milletari et al., 2016; Drozdzal et al., 2016) and cross-entropy loss instead of exclusively employing cross-entropy loss. Our experimental results demonstrate that the former yields better outcomes.

4.3   Demonstrated Task: `ShapeComposition`

To evaluate the effectiveness of our proposed redundancy filtering, we generated an artificial dataset containing 1000 sets of images. Each set consists of a triangle and an ellipse, deliberately overlapping in a specific region. The aim of the task is to take the input set of images and generate their union as the output, as illustrated in Figure 9. We employed the task decomposition approach in the CIML framework and assigned one of the figures as the primary modality and the other as the auxiliary modality. Note that both triangles and ellipses can be considered the primary modality.

In the implementation, we make several simplifications to the network architecture. Only one sub-network corresponds to the segmentation, and the other sub-network is utilized to acquire complementary information. Specifically, two encoders are employed independently to extract from the primary and auxiliary modalities, respectively. Two decoding pathways are then used. In the first path, one decoder inputs primary modality features (without gradient backpropagation) and auxiliary modality features, outputting μ and σ. 3D convolution layers without CIG modules are utilized to fuse features, thus eliminating the effect of CIG modules to verify the efficacy of our complementary information learning. Then, the reparameterization trick is used to sample complementary information features. In the second path, complementary information features are combined with the features directly extracted from the primary modality to predict the final results. Furthermore, as described in Section 3.2, the Kullback–Leibler divergence between the complementary information features and the standard normal distribution is minimized to constrain the complementary information features containing less information from the primary modality.

For qualitative analysis, we maintain the dimension of the complementary information features in line with the original image. As depicted in Figure 9, our proposed redundancy filtering-based complementary information extraction is effective, and the network efficiently extracts information from the auxiliary modality that is not present in the primary modality. Additionally, it contains little information that is already included in the primary modality.

4.4   Standardized Benchmarks

In this section, we compared our proposed CIML algorithms to eight state-of-the-art segmentation methods, including nnUNet (Isensee et al., 2018), AttentionUNet (Oktay et al., 2018), UNETR (Hatamizadeh et al., 2022), MAML (Zhang et al., 2021a), DIGEST (Li et al., 2022), RFNet (Ding et al., 2021), ACMINet (Zhuang et al., 2022) and NestedFormer (Xing et al., 2022). The first three methods are general methods, while the last five methods are designed specifically for multimodal segmentation. The nnUNet is a widely used benchmark that simplifies the critical decisions in designing an effective segmentation pipeline for any dataset. It and its variant, AttentionUNet, both employ early fusion. UNETR employs

Table 2: Quantitative comparison with SOTA methods on BraTS2020. WT, TC and ET signify the Dice coefficients of the whole tumor, the tumor core and the enhancing tumor, respectively. We use the results of RFNet and DIGEST from their paper. Other results are the result of reproducing them.

Methods Type	Methods	Patch Size	Dice↑ %				HD95↓			
			WT	TC	ET	MEAN	WT	TC	ET	MEAN
General Methods	nnUNet (Isensee et al., 2018)	128^3	91.59	87.50	83.47	87.52	4.75	4.0 8	5.60	4.81
	AttentionUNet (Oktay et al., 2018)	128^3	90.62	86.33	81.32	86.09	5.42	7.88	8.98	7.43
	UNETR (Hatamizadeh et al., 2022)	128^3	91.11	86.42	82.96	86.76	7.97	5.16	5.90	6.34
Multi-modal Methods	MAML (Zhang et al., 2021a)	128^3	91.40	88.05	82.40	87.28	4.84	5.95	7.90	7.82
	DIGEST (Li et al., 2022)	128^3	90.20	87.00	81.20	86.17	/	/	/	/
	RFNet (Ding et al., 2021)	128^3	91.11	85.21	78.00	84.77	/	/	/	/
	ACMINet (Zhuang et al., 2022)	128^3	91.79	87.99	82.56	87.45	5.70	5.09	6.95	5.91
	NestedFormer (Xing et al., 2022)	128^3	91.76	88.20	83.19	87.72	5.35	5.07	7.16	5.86
	CIML(Ours)	64^3	91.60	**89.14**	83.91	88.21	5.83	**3.70**	**4.95**	4.83
	CIML(Ours)	128^3	**91.88**	88.69	**84.34**	**88.30**	**4.58**	4.12	5.23	**4.64**

Table 3: Quantitative comparison with SOTA methods on the autoPET challenge. All outcomes of other algorithms result from reproducing them.

Methods Type	Methods	Dice ↑ %	HD95 ↓
General Methods	nnUNet (Isensee et al., 2018)	55.23	139.84
	AttentionUNet (Oktay et al., 2018)	57.87	108.62
	UNETR (Hatamizadeh et al., 2022)	41.62	177.04
Multi-modal Methods	MAML (Zhang et al., 2021a)	53.9	194.36
	ACMINet (Zhuang et al., 2022)	45.67	121.81
	NestedFormer (Xing et al., 2022)	52.51	183.14
	CIML(Ours)	**61.37**	**107.58**

Table 4: Quantitative comparison with SOTA methods on MICCAI HECKTOR 2022. All outcomes of other algorithms resulting from reproducing them.

Methods Type	Methods	Dice ↑ %	HD95 ↓
General Methods	nnUNet (Isensee et al., 2018)	73.61	6.09
	AttentionUNet (Oktay et al., 2018)	72.30	16.13
	UNETR (Hatamizadeh et al., 2022)	74.50	5.85
Multi-modal Methods	MAML (Zhang et al., 2021a)	70.00	29.05
	ACMINet (Zhuang et al., 2022)	74.23	8.13
	NestedFormer (Xing et al., 2022)	75.16	6.09
	CIML(Ours)	**76.28**	**5.17**

(a) Primary modality.   (b) Auxiliary modality.   (c) Predicted Segmentation.   (d) Ground Truth.   (e) Complementary Information.
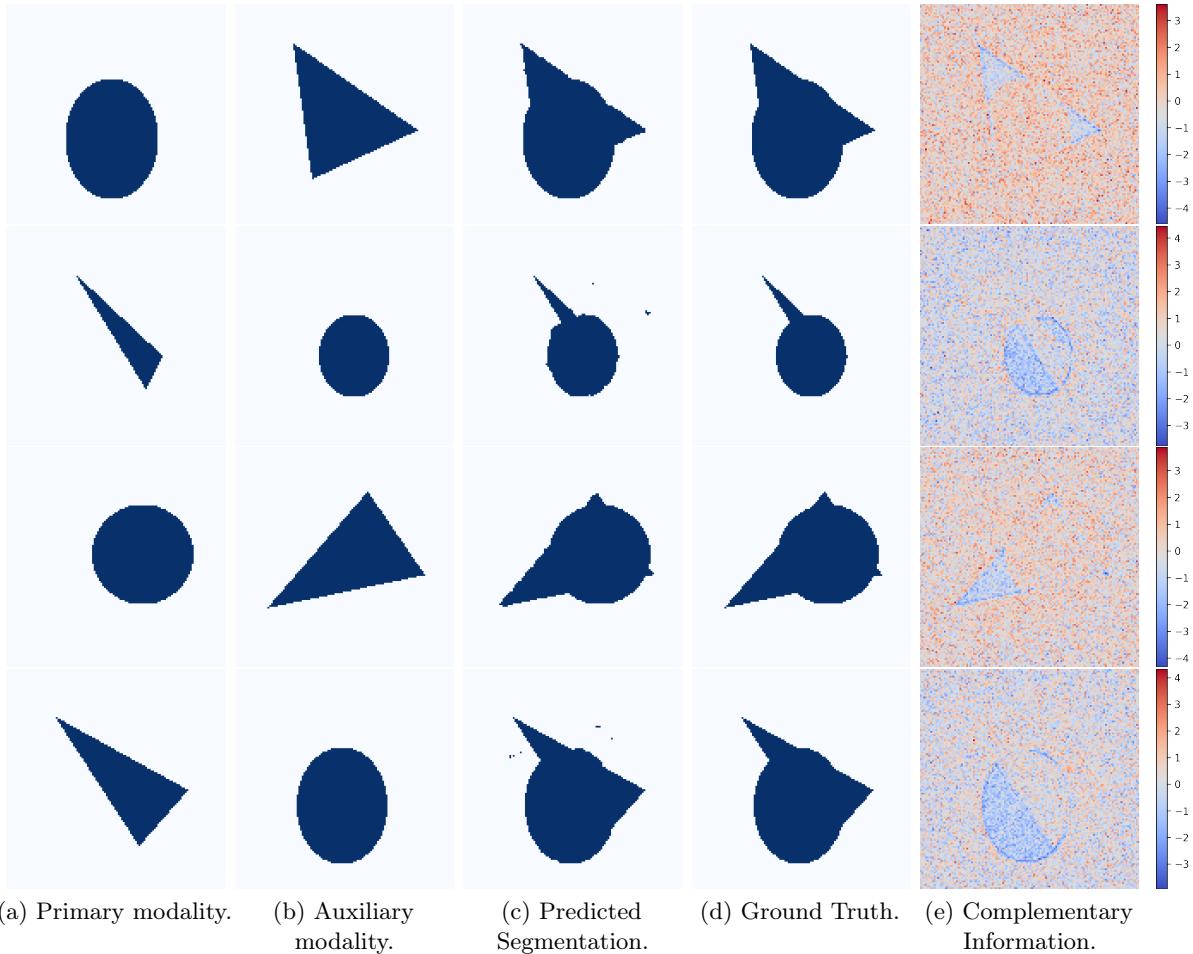
Figure 9: Each row displays the original images and results from a single case. The first and second columns depict the primary and auxiliary modalities, respectively, with the first and second images in the set. The third column shows the predicted segmentation, while the fourth column presents the ground-truth segmentation. The fifth column displays the visualization of the complementary information. Note: the images are best viewed in color for optimal clarity.

a transformer as the encoder to learn sequence representations of input images. MAML utilizes modality-specific encoders and incorporates a cross-modality attention mechanism for information fusion. DIGEST is a method applying a deeply supervised knowledge transfer network learning. RFNet also uses specific encoders like MAML and includes a region-aware module. ACMINet proposes a volumetric feature alignment module to align early features with late features. NestedFormer is a transformer-based method that designs a nested modality-aware feature aggregation module to model intra- and intermodality features for multimodal fusion.

Since RFNet and DIGEST are not open-source algorithms, we only compare CIML with them in the BraTS2020 dataset and offer the results from the original studies. Precisely,

we reproduce other methods using open-sourced codes. For a fair comparison, we employ the same training set, i.e., the same batch size, training epochs, and learning rate decay mechanism for all approaches.

As reported in Table 2, 3, and 4, our proposed method demonstrates superior performance compared to other methods, achieving the highest dice score and HD95 score in all regions across the three challenges. For the BraTS2020 dataset, we investigate the optimal performance of all models by using a patch size of $128 \times 128 \times 128$. Even with a smaller patch size, which means a more limited field of view, our method still outperforms most existing methods. Our method achieves higher segmentation accuracy compared to the state-of-the-art (SOTA) method, indicating that our algorithm can effectively eliminate the negative impact of *inter-modal* redundant information.

4.5 Ablation Study

Table 5: Ablation experiments on various components. The average Dice scores and HD95 scores are reported. "Message": models transport embeddings as messages; "Attention": cross-modal attention mechanism; "VI": conditional mutual information constrain;

Methods	Dice ↑%				HD95 ↓			
	WT	TC	ET	MEAN	WT	TC	ET	MEAN
Baseline	88.08	86.54	83.78	86.13	13.13	4.70	5.31	7.71
+ Message	90.30	88.37	81.38	86.68	8.71	6.22	6.63	7.18
+ Message + Attention	90.43	87.23	82.93	86.86	6.47	5.26	7.49	6.08
+ Message + VI	91.54	88.70	83.72	87.99	**4.60**	3.96	4.97	**4.51**
+ Message + Attention + VI	**91.60**	**89.14**	**83.91**	**88.21**	5.83	**3.70**	**4.95**	4.83

Table 6: Ablation study on various task decomposition. The average Dice scores and HD95 scores are reported. The results are listed from smallest to largest, according to the MEAN Dice.

Assignment				Dice ↑%				HD95 ↓			
FLAIR	T1	T2	T1CE	WT	TC	ET	MEAN	WT	TC	ET	MEAN
TC	TC, ET	WT	TC	91.01	88.95	81.01	86.99	6.01	4.55	6.85	5.80
WT	WT	TC	TC, ET	91.13	88.96	82.38	87.49	5.16	5.34	5.13	5.21
WT	WT, TC	TC	ET	91.21	89.12	82.41	87.58	**5.08**	6.18	5.24	5.50s
WT	WT, TC	WT, TC	TC, ET	**91.62**	88.20	83.79	87.87	5.98	3.97	5.17	5.04
WT	TC	WT, TC	TC, ET	91.60	**89.14**	**83.91**	**88.21**	5.83	**3.70**	**4.95**	**4.83**

4.5.1 Importance of Different Components

The efficacy of the proposed CIML segmentation method relies on several crucial components. An ablation study is conducted to evaluate the importance of each component and validate the efficiency of redundancy elimination, as shown in Table 5. The "Baseline" approach utilizes the nnUNet model for segmentation for each segmentor without incorporating information transport, resulting in each sub-model only being able to extract local information. The "+Message" notation represents the message passing between sub-models, where the local

CHUYUN SHEN ET AL.

embeddings are used as the message and fed into a one-layer 3D convolutional network equipped with a Batch Normalization layer and a sigmoid activation function, serving as a basic fusion module. The "+Attention" notation signifies the utilization of cross-model spatial attention mechanisms for integrating relevant information from the messages into the local embeddings. Finally, the "+VI" symbol represents the implementation of the variational inference for complementary information learning to extract complementary information from the messages.

Table 5 shows that the proposed message passing, cross-modal spatial attention mechanism, and complementary information learning significantly improved the model's performance. Specifically, the "baseline" approach that only utilizes local observations performs the worst in terms of the MEAN dice score. Message passing enhances the dice score on the WT and TC regions, as the input to the model contains complete information. Conversely, for the ET region, the opposite effect is observed. This is because the T1CE modality primarily determines the boundary of the ET region, and the features of other modalities do not contribute to determining the boundary of the ET region. This supports the rationale for the task decomposition in our framework. Additionally, the utilization of the cross-model spatial attention mechanism and our proposed redundancy filtering can significantly enhance the ability to extract relevant information from messages, resulting in better performance in the dice score and HD95. Notably, the implementation of our proposed redundancy filtering results in an even greater improvement. Combining the cross-modal spatial attention mechanism with the redundancy filtering yields optimal results.

4.5.2 Comparison of Different Task Decomposition

Based on our proposed CIML framework, the task decomposition is flexible. Different modalities have different clinical implications. On brain tumour segmentation, ET regions can be clearly discriminated on T1CE, and FLAIR is easier to discriminate WT regions, which have a clearer correspondence with the corresponding regions; TC regions have a relatively vague correspondence, so we have conducted some comparison experiments to test which assignment is better. We set 5 different assignment ways, and present results in Table 6. The results are ranked by MEAN Dice score, and the last assignment has the best MEAN Dice score. The first assignment does not assign the ET region to the T1CE modality and the WT region to the FLAIR modality, and the segmentation result is the worst, confirming the correlation between the modality and the target region and that it is reasonable for CIML to introduce human priori knowledge. Other assignments have similar segmentation results that exceed or are comparable to the results of nnUNet.

4.5.3 Hyperparametric Studies

In our work, there are two essential hyperparameters, the base number of filters and the hyperparameter β, which controls the tradeoff between the CE loss and KL loss.

As shown in Fig. 10, we explore the results of dice scores with filters of 4, 8, 16, 24 and 32. As base number filters increase, the mean dice (red dash line) score increases, and 24 and 32 base filters approach the best dice score.
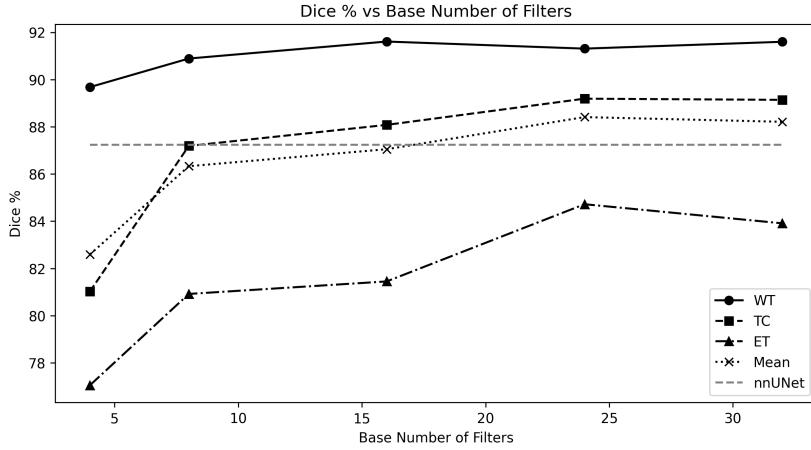
26

Figure 10: Visualization of dice score on the `BraTS2020` dataset for the different base numbers of filters, more filters mean wider network. The grey dash horizontal line indicates the average dice score of nnUNet.
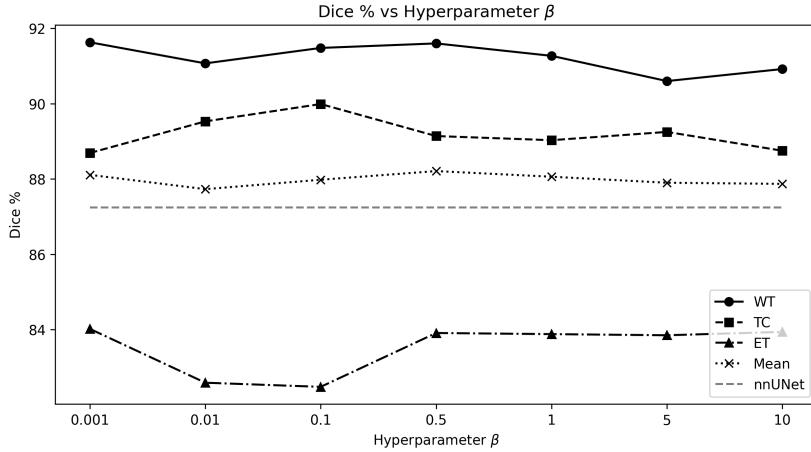

Figure 11: Visualization of dice score for the different hyperparameter β on the `BraTS2020` dataset. The grey dash horizontal line indicates the average dice score of nnUNet.

Additionally, we experiment with β from 1×10^{-3} to 10 and find that the highest average dice were achieved with β equals 0.5. All beta settings exceeded the results of nnUNet, indicating that our algorithm CIML is insensitive to β and has good generalization.

4.6 Visualization with Interpretability

4.6.1 Segmentation Results

As illustrated in Figure 13, we present the segmentation results of CIML alongside other state-of-the-art methods. Both nnUNet and our proposed method exhibit high accuracy across various cases. Notably, only UNETR and our method correctly classify the corresponding region as TC.
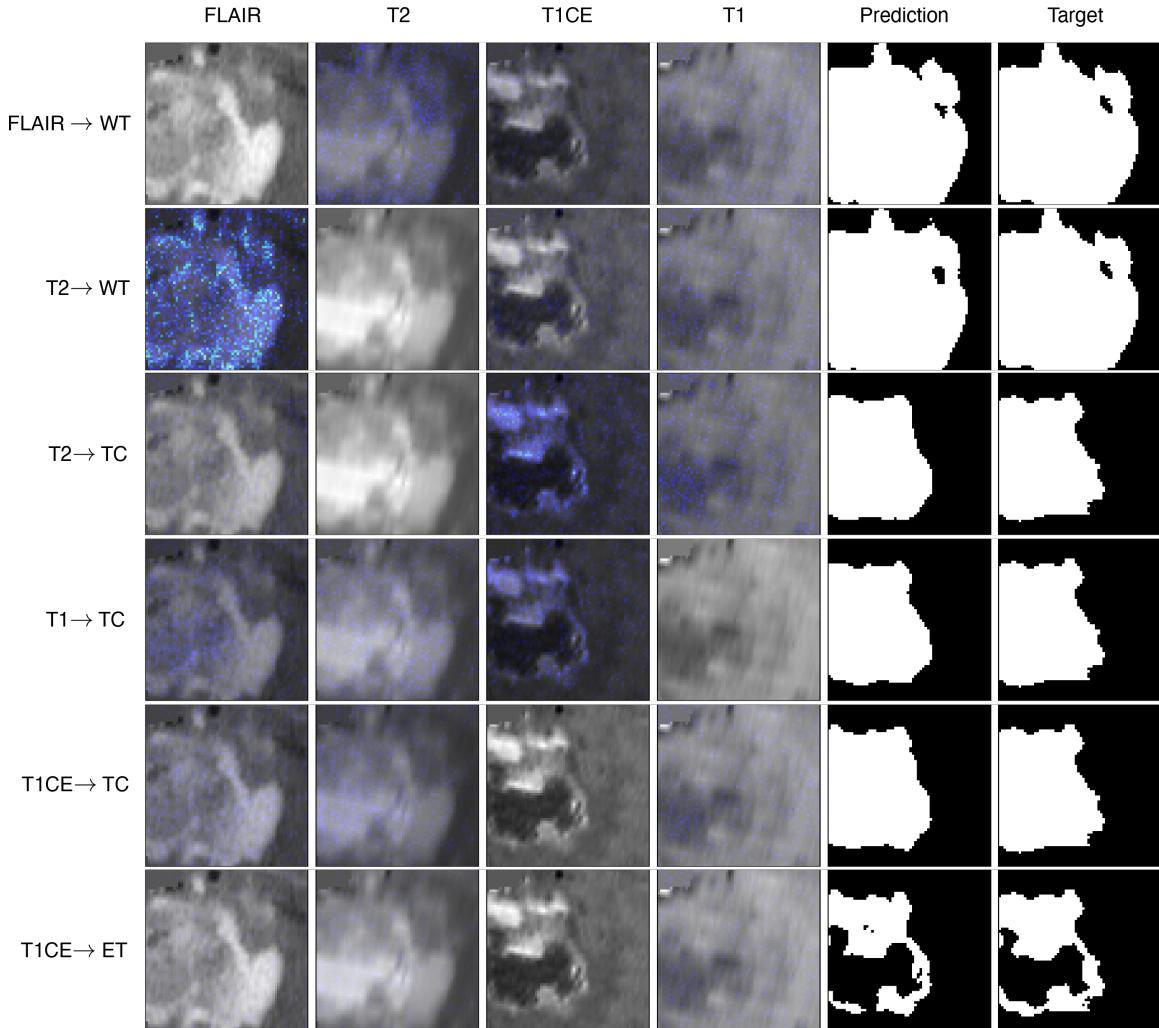
Figure 12: Information representation visualization of a case from the BraTS2020 challenge. Each row corresponds to a primary modality and a corresponding target region pair. The first four columns display the FLAIR, T2, T1CE, and T1 images and the information visualization representation masks (Dark blue indicates the smallest value, light blue means the middle value, and yellow illustrates the largest value). Specifically, the primary modal displays only the original image without masks. The fifth column, 'Prediction,' displays prediction outcomes, while the final column, 'Target,' displays the ground truth. Note: the images are best viewed in color for optimal clarity.
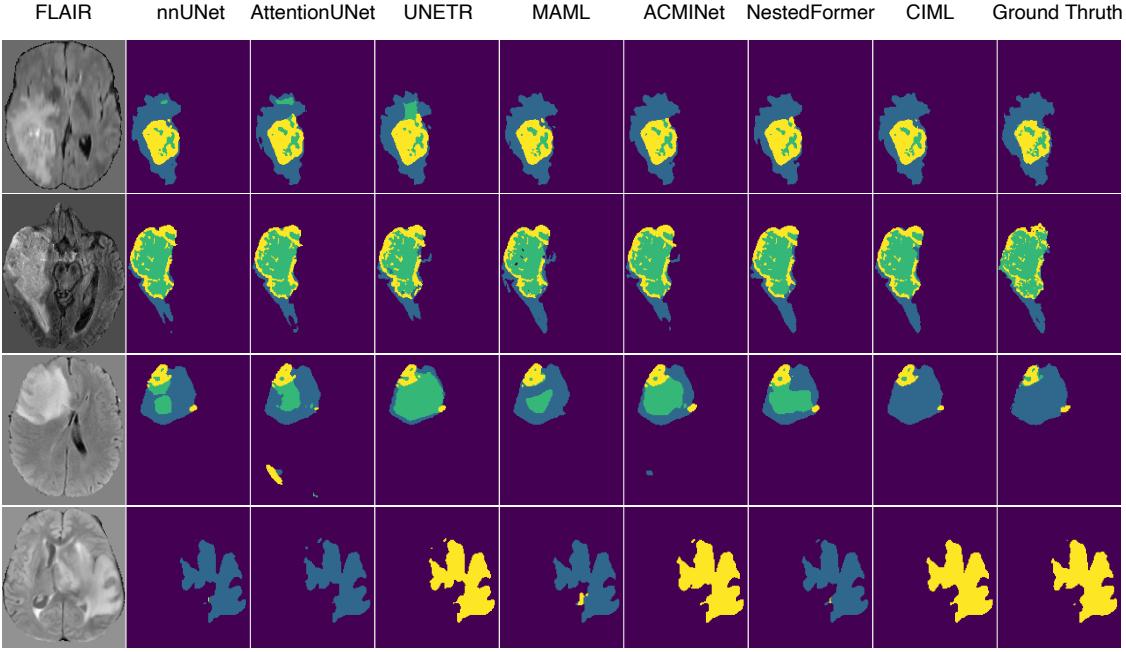
Figure 13: Visualization of the predicted segmentation maps compared with the SOTA methods. The first column shows FLAIR images, and the other column shows the prediction results and the corresponding ground truth. Note: the images are best viewed in color for optimal clarity.

4.6.2 COMPLEMENTARY INFORMATION

In subsection 4.3, the demonstration experiment verifies that our proposed redundancy filtering is efficient. In CIML, the extracted information representations are high-dimensional features, and to visualize the complementary information, we use the Grad-CAM algorithm to get class-discriminative localization maps (we will refer to it as a heatmap when there is no possibility of confusion) that highlight the voxels focused on by our proposed CIML algorithm.

In deep convolutional networks, the deeper layers typically extract semantic information but lose positional correspondence. Conversely, the shallower layers are able to extract detailed pattern information while preserving positional correspondence. Therefore, in order to maintain a clear understanding of positional information, we choose to visualize the information representations from the shallowest layers of the network, which have the same resolutions as the images with C channels. In the `BraTS2020` dataset, we default decompose segmentation into four segmentors. Additionally, there are six modality target region pairs. For i-th pair, let $\{A_i\}_{c=1}^{C}$ be the information representations (with C channels), and y^k be the logit for a chosen pixel class k. Grad-CAM averages the partial gradients of y^k with respect to N voxels of each information representation. The heatmap for class k in pair i

$$\zeta_i^k = \text{ReLU}\left(\sum_c \alpha_k^c A_i^c\right), \tag{14}$$

with

$$\alpha_k^c = \frac{1}{N} \sum_n \frac{\partial y^k}{\partial A_{i,n}^c}, \tag{15}$$

where α_k^c is the neuron importance weights of the $c-th$ channel of information representations for pair i.

In Figure 12, we visualize information representations extracted from the last messages by applying heatmaps as masks added to the original images. The last message refers to the shallow embedding in the neural network, which is the message with the highest resolution. We normalize the heatmaps across all rows so that the amount of information in each auxiliary modality can be easily compared. Yellow represents the largest value, dark blue represents the smallest value, and light blue represents the middle value. As shown in the first row, FLAIR is the primary modality, and the T2 image contains the most complementary information compared to the other two modalities. In the second row, T2 is the primary modality, and the FLAIR image contains the most complementary information. Additionally, the left-down region of the FLAIR image contains more information, which is consistent with medical domain knowledge. This region, depicted in hyperintensity (lighter in source images), indicates the presence of edema, typically locates at the periphery of the WT. The third and fourth rows illustrate TC is the target region and the hyperintensity regions in T1CE, which means ET regions, contain the most information. In the final two rows concerning T1CE as the primary modality, it is observed that auxiliary modalities contribute less information overall. However, the T2 image still provides some valuable complementary insights, particularly in the hyperintense areas, which appear as low-intensity zones in the T1CE image indicating the necrotic and non-enhancing tumor core. These results demonstrate that the results predicted by our proposed CIML methods are consistent with the physician's domain knowledge and permit further verification that the algorithm can extract complementary information from high-dimensional data.

4.6.3 COMPLEMENTARY INFORMATION WEIGHTS

As shown in Figure 12, auxiliary modalities contain various complementary information for each *modality-region* pair. To further explore the contribution of different auxiliary modalities to segmentation, we propose using the complementary information weight to quantify the contribution. We define the complementary information weight for pair i as

$$\varpi_i^k = \frac{\sum_n \zeta_{i,n}^k}{\sum_j \sum_n \zeta_{j,n}^k}. \tag{16}$$

Figure 14 reveals that FLAIR provides the most substantial complementary information weight for the T2 segmenting of the WT target region. Additionally, when segmenting the TC and ET target regions, T1CE requires minimal supplementary information. Furthermore, T1CE offers more complementary information than other modalities, especially for subtasks where the TC target region is segmented using T1 and T2. These findings are consistent with the observations in Figure 2, where T1CE is sensitive to NCR/NET and ET regions, while FLAIR is sensitive to the WT target region. Thus, CIML can effectively prioritize sensitive modalities and extract discriminative features from auxiliary modalities.
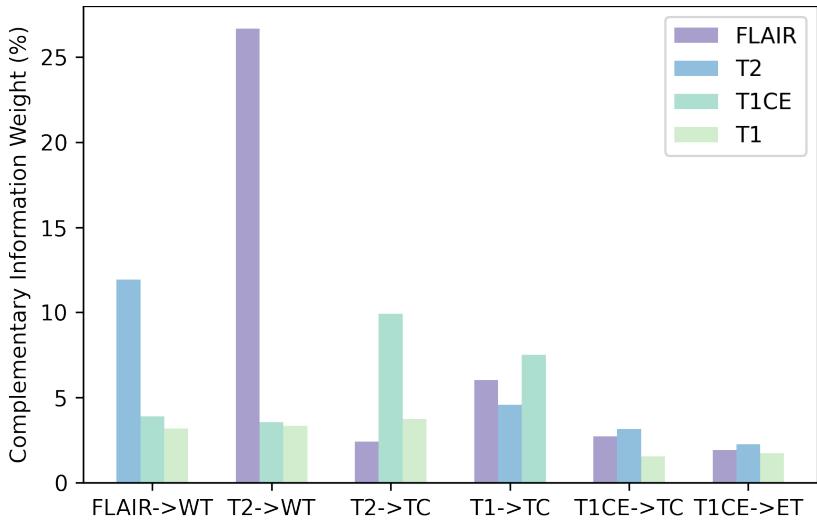
Figure 14: Visualization of the complementary information weights by our CIML on the `BraTS2020` dataset.

5. Acknowledgements

6. Conclusion

In this study, we propose the **complementary information mutual learning (CIML)** framework, which provides a unique solution to the problem of *inter-modal* redundant information in multimodal learning, an issue not addressed by previous state-of-the-art (SOTA) methods. Our framework, based on *addition* operation, provides a systematic approach by employing inductive bias for task decomposition and message passing for redundancy filtering, thereby enhancing the effectiveness of multimodal medical image segmentation. We extensively evaluate our approach and demonstrate its effectiveness, outperforming current SOTA methods. Furthermore, the message passed through redundancy filtering enables the application of visualization techniques such as Grad-CAM, thus improving the interpretability of the algorithm. In conclusion, our proposed CIML framework has the potential to significantly enhance the quality and reliability of multimodal medical image segmentation, ultimately leading to improved clinical diagnosis and treatment outcomes.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, pages 1–32, 2021.

Avi Ben-Cohen, Eyal Klang, Stephen P Raskin, Shelly Soffer, Simona Ben-Haim, Eli Konen, Michal Marianne Amitai, and Hayit Greenspan. Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. *Engineering Applications of Artificial Intelligence*, 78:186–194, 2019.

Piet M Bouman, Samantha Noteboom, Fernando A Nobrega Santos, Erin S Beck, Gregory Bliault, Marco Castellaro, Massimiliano Calabrese, Declan T Chard, Paul Eichinger, Massimo Filippi, et al. Multicenter evaluation of ai-generated dir and psir for cortical and juxtacortical multiple sclerosis lesion detection. *Radiology*, page 221425, 2023.

Yujia Cao, Mariët Theune, and Anton Nijholt. Modality effects on cognitive load and performance in high-load information presentation. In *IUI*, 2009.

Leonardo G Cohen, Pablo Celnik, Alvaro Pascual-Leone, Brian Corwell, Lala Faiz, James Dambrosia, Manabu Honda, Norihiro Sadato, Christian Gerloff, M Dolores Catalá, et al. Functional relevance of cross-modal plasticity in blind humans. *Nature*, 389(6647):180–183, 1997.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *ICCV*, 2021.

Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.

Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 74–82. Springer, 2019.

Michal Drozdzal, Eugene Vorontsov, Gabriel Chartrand, Samuel Kadoury, and Chris Pal. The importance of skip connections in biomedical image segmentation. In *DLMIA*, 2016.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):1–7, 2022.

Michael S Gazzaniga, Richard B Ivry, and GR Mangun. Cognitive neuroscience. the biology of the mind,(2014), 2006.

Steven M Greenberg, Wendy C Ziai, Charlotte Cordonnier, Dar Dowlatshahi, Brandon Francis, Joshua N Goldstein, J Claude Hemphill III, Ronda Johnson, Kiffon M Keigher, William J Mack, et al. 2022 guideline for the management of patients with spontaneous intracerebral hemorrhage: a guideline from the american heart association/american stroke association. *Stroke*, 53(7):e282–e361, 2022.

Ziqin Han, Qiuying Chen, Lu Zhang, Xiaokai Mo, Jingjing You, Luyan Chen, Jin Fang, Fei Wang, Zhe Jin, Shuixing Zhang, et al. Radiogenomic association between the t2-flair mismatch sign and idh mutation status in adult patients with lower-grade gliomas: An updated systematic review and meta-analysis. *European Radiology*, 32(8):5339–5352, 2022.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation. In *WACV*, 2022.

Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. Hemis: Hetero-modal image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 469–477. Springer, 2016.

Jeff Henrikson. Completeness and total boundedness of the hausdorff metric. *MIT Undergraduate Journal of Mathematics*, 1(69-80):10, 1999.

Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in Medicine & Biology*, 46(3):R1, 2001.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Zhenbo Huang, Shiliang Sun, Jing Zhao, and Liang Mao. Multi-modal policy fusion for end-to-end autonomous driving. *Information Fusion*, 98:101834, 2023.

Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, et al. NNU-Net: Self-adapting framework for U-net-based medical image segmentation. *ArXiv preprint*, abs/1809.10486, 2018. URL https://arxiv.org/abs/1809.10486.

Carolien AN Knoop-van Campen, Eliane Segers, and Ludo Verhoeven. Modality and redundancy effects, and their relation to executive functioning in children with dyslexia. *Research in Developmental Disabilities*, 90:41–50, 2019.

Changhee Lee and Mihaela van der Schaar. A variational information bottleneck approach to multi-omics data integration. In *AISTATS*, 2021.

Chao Li, Shuo Wang, Pan Liu, Turid Torheim, Natalie R Boonzaier, Bart RJ van Dijken, Carola-Bibiane Schönlieb, Florian Markowetz, and Stephen J Price. Decoding the interdependence of multiparametric magnetic resonance imaging to reveal patient subgroups correlated with survivals. *Neoplasia*, 21(5):442–449, 2019.

Chao Li, Wenjian Huang, Xi Chen, Yiran Wei, Lipei Zhang, Jianguo Zhang, Stephen Price, and Carola-Bibiane Schönlieb. Expectation-maximization regularised deep learning for tumour segmentation. In *2023 IEEE 20th international symposium on biomedical imaging (ISBI)*, pages 1–5. IEEE, 2023.

Haoran Li, Cheng Li, Weijian Huang, Xiawu Zheng, Yan Xi, and Shanshan Wang. Digest: Deeply supervised knowledge transfer network learning for brain tumor segmentation with incomplete multi-modal mri scans. *arXiv preprint arXiv:2211.07993*, 2022.

Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng. Adversarial multimodal representation learning for click-through rate prediction. In *WWW*, 2020.

Xuelong Li. Multi-modal cognitive computing. *SCIENTIA SINICA Informationis*, 53(1): 1–32, 2023.

Xuelong Li, Dacheng Tao, Stephen J Maybank, and Yuan Yuan. Visual music and musical vision. *Neurocomputing*, 71(10-12):2023–2028, 2008.

Wen-Wei Lin, Cheng Juang, Mei-Heng Yueh, Tsung-Ming Huang, Tiexiang Li, Sheng Wang, and Shing-Tung Yau. 3d brain tumor segmentation using a two-stage optimal mass transport algorithm. *Scientific Reports*, 11(1):14686, 2021.

Yuhan Liu, Minzhi Yin, and Shiliang Sun. Detexnet: accurately diagnosing frequent and challenging pediatric malignant tumors. *IEEE Transactions on Medical Imaging*, 40(1): 395–404, 2020.

Zecheng Liu, Jia Wei, Rui Li, and Jianlong Zhou. Sfusion: Self-attention based n-to-one multimodal fusion block. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 159–169. Springer, 2023.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 2022.

Richard E Mayer and Roxana Moreno. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1):43–52, 2003.

John McCarthy, Marvin L Minsky, Nathaniel Rochester, and Claude E Shannon. A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Magazine*, 27(4):12–12, 2006.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *ArXiv preprint*, abs/1804.03999, 2018. URL https://arxiv.org/abs/1804.03999.

Mauricio Orbes-Arteaga, M Jorge Cardoso, Lauge Sørensen, Marc Modat, Sébastien Ourselin, Mads Nielsen, and Akshay Pai. Simultaneous synthesis of flair and segmentation of white matter hypointensities from t1 mris. *arXiv preprint arXiv:1808.06519*, 2018.

Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallières, Simeng Zhu, Juanying Xie, Ying Peng, et al. Head and neck tumor segmentation in pet/ct: the hecktor challenge. *Medical Image Analysis*, 77: 102336, 2022.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, Tom Eccles, Jake Bruce, Ali Razavi, Ashley Edwards, Nicolas Heess, Yutian Chen, Raia Hadsell, Oriol Vinyals, Mahyar Bordbar, and Nando de Freitas. A generalist agent. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=1ikK0kHjvj. Featured Certification.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *ITW*, 2015.

Christopher Tosh, Akshay Krishnamurthy, and Daniel Hsu. Contrastive learning, multi-view redundancy, and linear models. In *Algorithmic Learning Theory*, pages 1179–1206. PMLR, 2021.

Gijs Van Tulder and Marleen de Bruijne. Why does synthesized data improve multi-sequence classification? In *MICCAI*, 2015.

Kira Vinogradova, Alexandr Dibrov, and Gene Myers. Towards interpretable semantic segmentation via gradient-weighted class activation mapping (student abstract). In *AAAI*, 2020.

Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023a.

Jingyao Wang, Luntian Mou, Lei Ma, Tiejun Huang, and Wen Gao. Amsa: Adaptive multimodal learning for sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(3s), 2022.

Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. Deep multi-view information bottleneck. In *ICDM*, 2019.

Shuai Wang, Zipei Yan, Daoan Zhang, Haining Wei, Zhongsen Li, and Rui Li. Prototype knowledge distillation for medical segmentation with missing modality. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023b.

Yiran Wei, Xi Chen, Lei Zhu, Lipei Zhang, Carola-Bibiane Schönlieb, Stephen Price, and Chao Li. Multi-modal learning for predicting the genotype of glioma. *IEEE Transactions on Medical Imaging*, 42(11):3167–3178, 2023.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *NeurIPS*, 2018.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *MICCAI*, 2022.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *arXiv preprint arXiv:2206.06488*, 2022.

Wenfang Yao, Kejing Yin, William K Cheung, Jia Liu, and Jing Qin. Drfuse: Learning disentangled representation for clinical multi-modal fusion with missing modality and modal inconsistency. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16416–16424, 2024.

Ben P Yuhas, Moise H Goldstein, and Terrence J Sejnowski. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, 27(11): 65–71, 1989.

Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2418–2428, 2022.

Xilin Zhang, Li Zhaoping, Tiangang Zhou, and Fang Fang. Neural activities in v1 create a bottom-up saliency map. *Neuron*, 73(1):183–192, 2012.

Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *MICCAI*, 2021a.

Yue Zhang, Pinyuan Zhong, Dabin Jie, Jiewei Wu, Shanmei Zeng, Jianping Chu, Yilong Liu, Ed X Wu, and Xiaoying Tang. Brain tumor segmentation from multi-modal mr images via ensembling unets. *Frontiers in Radiology*, page 11, 2021b.

Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Comput. Surv.*, 56(9), apr 2024. ISSN 0360-0300. doi: 10.1145/3649447. URL https://doi.org/10.1145/3649447.

Liang Zhao, Tao Yang, Jie Zhang, Zhikui Chen, Yi Yang, and Z Jane Wang. Co-learning non-negative correlated and uncorrelated features for multi-view data. *IEEE Transactions on Neural Networks and Learning Systems*, 32(4):1486–1496, 2020.

Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.

Chenhong Zhou, Changxing Ding, Xinchao Wang, Zhentai Lu, and Dacheng Tao. One-pass multi-task networks with cross-task guided attention for brain tumor segmentation. *IEEE Transactions on Image Processing*, 29:4516–4529, 2020a.

Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE Transactions on Medical Imaging*, 39 (9):2772–2781, 2020b.

Tongxue Zhou, Su Ruan, and Stéphane Canu. A review: Deep learning for medical image segmentation using multi-modality fusion. *Array*, 3:100004, 2019.

Tongxue Zhou, Su Ruan, Pierre Vera, and Stéphane Canu. A tri-attention fusion guided multi-modal segmentation network. *Pattern Recognition*, 124:108417, 2022.

Yaochen Zhu, Jiayi Xie, and Zhenzhong Chen. Predicting the popularity of micro-videos with multimodal variational encoder-decoder framework. *arXiv preprint arXiv:2003.12724*, 2020.

Yuzhou Zhuang, Hong Liu, Enmin Song, and Chih-Cheng Hung. A 3d cross-modality feature interaction network with volumetric feature alignment for brain tumor and tissue segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2022.

Appendix A. Mutual Information and Venn Diagrams

A.1 Entropy

Entropy is a fundamental concept in information theory, quantifying the uncertainty or randomness in a random variable. Formally, the entropy $H(X)$ of a continuous random variable X with probability density function $p(x)$ is defined as:

$$H(X) = -\int dx\, dy p(x) \log p(x), \tag{17}$$

where $p(x)$ is the probability density of x. Entropy measures the expected amount of information required to describe the outcome of X.

A.2 Mutual Information

Mutual information (MI) quantifies the amount of information one random variable contains about another. For two continuous random variables X and Y with a joint probability density function $p(x, y)$, the mutual information $\mathcal{I}(X; Y)$ is defined as:

$$\mathcal{I}(X; Y) = \int dx\, dy p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \tag{18}$$

where $p(x)$ and $p(y)$ are the marginal probability density functions of X and Y, respectively. MI quantifies the reduction in uncertainty about X given knowledge of Y.

A.3 Conditional Mutual Information

Conditional mutual information (CMI) extends MI to account for a third random variable Z. The conditional mutual information $\mathcal{I}(X; Y|Z)$ is defined as:

$$\mathcal{I}(X; Y|Z) = \int dx\, dy\, dz p(z)p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}, \tag{19}$$

where $p(x, y|z)$ is the conditional joint probability density function of X and Y given Z, and $p(x|z)$ and $p(y|z)$ are the conditional marginal densities. CMI measures the information shared between X and Y accounting for the effect of Z.

A.4 Relationship with Venn Diagrams

Venn diagrams provide a visual representation to illustrate relationships between entropy, MI, and CMI. In a Venn diagram, the entropy $H(X)$ of a random variable X is depicted as the area of a circle, representing the total uncertainty of X. Similarly, the entropy $H(Y)$ of another random variable Y is represented by another circle.

The overlap between the two circles represents the mutual information $\mathcal{I}(X; Y)$, which quantifies the shared information between X and Y and the reduction in uncertainty of one variable due to knowledge of the other.

Introducing a third random variable Z, the conditional mutual information $\mathcal{I}(X;Y|Z)$ can be visualized as the overlap between the areas of X and Y, excluding the part influenced by Z. In a Venn diagram with three circles representing X, Y, and Z, CMI focuses on the shared area between X and Y while excluding the part of the circle Z.

Venn diagrams provide an intuitive method to understand interactions between different information-theoretic measures, highlighting dependencies and shared information among multiple random variables.

A.5   Common Formulas and Properties for Mutual Information

Symmetry: Mutual information is symmetric, which means that the mutual information between two random variables X and Y is the same regardless of the order in which the variables are considered. Mathematically, this can be expressed as:

$$\mathcal{I}(X;Y) = \mathcal{I}(Y;X). \tag{20}$$

This property follows from the definition of mutual information, which is based on joint and marginal probabilities that are symmetric with respect to X and Y.

Non-negativity: Mutual information is always non-negative. This means that the mutual information between any two random variables X and Y is greater than or equal to zero:

$$\mathcal{I}(X;Y) \geq 0. \tag{21}$$

This property arises because mutual information is a measure of the reduction in uncertainty about one random variable given knowledge of another, and this reduction in uncertainty cannot be negative.

Chain Rule: The chain rule for mutual information allows us to break down the mutual information between multiple variables into simpler components. For three random variables A, B, and C, the chain rule states:

$$\mathcal{I}(A,B;C) = \mathcal{I}(A;C) + \mathcal{I}(B;C \mid A). \tag{22}$$

This rule can be extended to more variables. For example, for four random variables A, B, C, and D, the chain rule becomes:

$$\mathcal{I}(A,B,C;D) = \mathcal{I}(A;D) + \mathcal{I}(B;D \mid A) + \mathcal{I}(C;D \mid A,B). \tag{23}$$

In general, for random variables $X_1, X_2, \ldots, X_n$ and Y, the chain rule is given by:

$$\mathcal{I}(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} \mathcal{I}(X_i; Y \mid X_1, X_2, \ldots, X_{i-1}). \tag{24}$$

This powerful rule helps in decomposing mutual information into more manageable parts, which can be particularly useful in proofs and derivations.

Appendix B. Bayesian Graph Representation

B.1   Introduction to Bayesian Graph Representation

Bayesian networks, also known as Bayesian graph representations, are powerful tools for representing and reasoning about probabilistic dependencies and causal relationships among a set of variables. A Bayesian network is a directed acyclic graph (DAG) where each node represents a random variable, and the edges represent probabilistic dependencies between these variables. Bayesian networks are particularly useful for causal inference because the directed edges can be interpreted as causal influences. If there is a directed edge from X_i to X_j, it suggests that X_i is a direct cause of X_j. This causal interpretation allows for the modeling of complex causal relationships and the prediction of the effects of interventions.

B.2   Derivation for Equation (4)

We aim to derive the joint probability distribution:

$$p(X_1, X_2, Y_1, K_2) = p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2, Y_1). \tag{25}$$

Consider the Bayesian network depicted in Figure 15, which illustrates the relationships among the variables X, X_1, X_2, Y_1, and K_2.
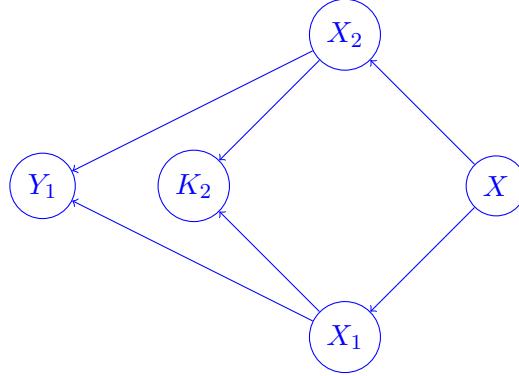

Figure 15: A Bayesian network representing the relationship of modalities X_1, X_2, target Y_1, and representation K_2.

From the graph, it is clear that Y_1 and K_2 have X_1 and X_2 as their parent nodes, and X_1 and X_2 have X as their parent node. The joint distribution of the variables can be expressed as:

$$p(X_1, X_2, Y_1, K_2, X) = p(Y_1 \mid X_1, X_2) \cdot p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2 \mid X) \cdot p(X). \tag{26}$$

To derive the desired joint distribution, we integrate over the latent variable X:

$$
\begin{aligned}
p(X_1, X_2, Y_1, K_2) &= \int p(X_1, X_2, Y_1, K_2, X) \, dX \\
&= \int p(Y_1 \mid X_1, X_2) \cdot p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2 \mid X) \cdot p(X) \, dX \\
&= \int p(Y_1 \mid X_1, X_2) \cdot p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2, X) \, dX \\
&= p(Y_1 \mid X_1, X_2) \cdot p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2) \\
&= p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2, Y_1).
\end{aligned}
\tag{27}
$$

Thus, we have derived that

$$
p(X_1, X_2, Y_1, K_2) = p(K_2 \mid X_1, X_2) \cdot p(X_1, X_2, Y_1),
\tag{28}
$$

as required.

Appendix C. Mutual Information Derivation

C.1 Derivation for Equation (2)

To prove the given equation,

$$
\mathcal{I}_1(X_1, X_2; K_2) = \underbrace{\mathcal{I}_2(K_2; Y_1 \mid X_1)}_{\text{complementary predictive information}} \\
+ \underbrace{\mathcal{I}_3(K_2; X_1)}_{\text{duplicated information}} \\
+ \underbrace{\mathcal{I}_4(K_2; X_2 \mid X_1, Y_1)}_{\text{unique but irrelevant information}},
\tag{29}
$$

we start by expressing each term and using the chain rule for mutual information.

First, using the chain rule for mutual information:

$$
\mathcal{I}_1(X_1, X_2; K_2) = \mathcal{I}_3(K_2; X_1) + \mathcal{I}(X_2; K_2 \mid X_1).
\tag{30}
$$

Next, we expand $\mathcal{I}(X_2; K_2 \mid X_1)$ by the chain rule for mutual information:

$$
\mathcal{I}(X_2; K_2 \mid X_1) = \mathcal{I}_2(K_2; Y_1 \mid X_1) + \mathcal{I}_4(K_2; X_2 \mid X_1, Y_1).
\tag{31}
$$

Combining the above results, we get:

$$
\mathcal{I}_1(X_1, X_2; K_2) = \underbrace{\mathcal{I}_2(K_2; Y_1 \mid X_1)}_{\text{complementary predictive information}} + \underbrace{\mathcal{I}_3(K_2; X_1)}_{\text{duplicated information}} + \underbrace{\mathcal{I}_4(K_2; X_2 \mid X_1, Y_1)}_{\text{unique but irrelevant information}}.
\tag{32}
$$

Thus, the equation is proved.

C.2 Derivation of Equation (5)

To prove the given equation:

$$
\begin{aligned}
\mathcal{I}_1(X_1, X_2; K_2) &= \int p(x_1, x_2, \kappa_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{p(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2 \\
&\leq \int p(x_1, x_2, \kappa_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{r(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2 \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \int p(\kappa_2 \mid x_1^i, x_2^i) \log\left(\frac{p(\kappa_2 \mid x_1^i, x_2^i)}{r(\kappa_2)}\right) d\kappa_2,
\end{aligned}
\tag{33}
$$

First, based on the definition of mutual information, we have:

$$
\mathcal{I}_1(X_1, X_2; K_2) = \int p(x_1, x_2, \kappa_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{p(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2,
\tag{34}
$$

Next, by substituting $p(x_1, x_2, \kappa_2)$, we obtain:

$$
\mathcal{I}_1(X_1, X_2; K_2) = \int p(\kappa_2 \mid x_1, x_2) p(x_1, x_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{p(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2.
\tag{35}
$$

To derive the inequality, we apply the non-negativity of the Kullback-Leibler (KL) divergence:

$$
D_{\mathrm{KL}}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \geq 0,
\tag{36}
$$

which implies:

$$
\int p(\kappa_2 \mid x_1, x_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{p(\kappa_2)}\right) d\kappa_2 \leq \int p(\kappa_2 \mid x_1, x_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{r(\kappa_2)}\right) d\kappa_2.
\tag{37}
$$

Thus, we have:

$$
\begin{aligned}
&\int p(\kappa_2 \mid x_1, x_2) p(x_1, x_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{p(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2 \\
&\leq \int p(\kappa_2 \mid x_1, x_2) p(x_1, x_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{r(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2.
\end{aligned}
\tag{38}
$$

Finally, using Monte Carlo sampling, we approximate:

$$
\begin{aligned}
&\int p(x_1, x_2, \kappa_2) \log\left(\frac{p(\kappa_2 \mid x_1, x_2)}{r(\kappa_2)}\right) dx_1\, dx_2\, d\kappa_2 \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \int p(\kappa_2 \mid x_1^i, x_2^i) \log\left(\frac{p(\kappa_2 \mid x_1^i, x_2^i)}{r(\kappa_2)}\right) d\kappa_2.
\end{aligned}
\tag{39}
$$

C.3 Derivation of Equation (6)

To prove the given equation:

$$
\begin{aligned}
&\mathcal{I}_2(K_2; Y_1 \mid X_1) \\
&= \int p(x_1, x_2, \kappa_2, y_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, dx_2\, d\kappa_2\, dy_1 \\
&\geq \int p(x_1, x_2, \kappa_2, y_1) \log\left(\frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, dx_2\, d\kappa_2\, dy_1 \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \int p(\kappa_2 \mid x_1^i, x_2^i) \log q(y_1^i \mid \kappa_2, x_1^i)\, d\kappa_2 + H.
\end{aligned}
\tag{40}
$$

First, based on the definition of mutual information, we get:

$$
\mathcal{I}_2(K_2; Y_1 \mid X_1) = \int p(\kappa_2, x_1, y_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, d\kappa_2\, dy_1.
\tag{41}
$$

Then, based on the integral, the joint distribution is equal to the marginal distribution:

$$
p(\kappa_2, x_1, y_1) = \int p(\kappa_2, x_1, x_2, y_1)\, dx_2,
\tag{42}
$$

we obtain:

$$
\mathcal{I}_2(K_2; Y_1 \mid X_1) = \int p(\kappa_2, x_1, x_2, y_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, dx_2\, d\kappa_2\, dy_1.
\tag{43}
$$

To derive the inequality, we apply the non-negativity of the Kullback-Leibler (KL) divergence:

$$
D_{\mathrm{KL}}(P\|Q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \geq 0,
\tag{44}
$$

which implies:

$$
\int p(y_1 \mid \kappa_2, x_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dy_1 \geq \int p(y_1 \mid \kappa_2, x_1) \log\left(\frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dy_1.
\tag{45}
$$

Then:

$$
\begin{aligned}
&\int p(\kappa_2, x_1)\, d\kappa_2\, dx_1 \int p(y_1 \mid \kappa_2, x_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dy_1 \\
&\geq \int p(\kappa_2, x_1)\, d\kappa_2\, dx_1 \int p(y_1 \mid \kappa_2, x_1) \log\left(\frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dy_1.
\end{aligned}
\tag{46}
$$

Thus:

$$
\begin{aligned}
&\int p(x_1, x_2, \kappa_2, y_1) \log\left(\frac{p(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, dx_2\, d\kappa_2\, dy_1 \\
&\geq \int p(x_1, x_2, \kappa_2, y_1) \log\left(\frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)}\right) dx_1\, dx_2\, d\kappa_2\, dy_1.
\end{aligned}
\tag{47}
$$

43

Further, we can split the upper bound into two terms:

$$\int p(x_1, x_2, \kappa_2, y_1) \log q(y_1 \mid \kappa_2, x_1) \, dx_1 \, dx_2 \, d\kappa_2 \, dy_1$$
$$= \int p(x_1, x_2, \kappa_2, y_1) \log q(y_1 \mid \kappa_2, x_1) \, dx_1 \, dx_2 \, d\kappa_2 \, dy_1 - H, \tag{48}$$

where

$$H = \int p(x_1, x_2, \kappa_2, y_1) \log p(y_1 \mid x_1) \, dx_1 \, dx_2 \, d\kappa_2 \, dy_1$$
$$= \int p(x_1, y_1) \log p(y_1 \mid x_1) \, dx_1 \, dy_1, \tag{49}$$

because H is independent of our optimization procedure and so can be ignored. For the approximation using Monte Carlo sampling:

$$\int p(x_1, x_2, \kappa_2, y_1) \log \left(\frac{q(y_1 \mid \kappa_2, x_1)}{p(y_1 \mid x_1)} \right) \, dx_1 \, dx_2 \, d\kappa_2 \, dy_1$$
$$\approx \frac{1}{N} \sum_{i=1}^{N} \int p(\kappa_2 \mid x_1^i, x_2^i) \log q(y_1^i \mid \kappa_2, x_1^i) \, d\kappa_2 + H. \tag{50}$$