PERSONALIZE TO GENERALIZE: TOWARDS A UNIVERSAL MEDICAL MULTI-MODALITY GENERALIZATION THROUGH PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Personalized medicine is a groundbreaking healthcare framework for the 21^{st} century, tailoring medical treatments to individuals based on unique clinical characteristics, including diverse medical imaging modalities. Given the significant differences among these modalities due to distinct underlying imaging principles, generalization in multi-modal medical image tasks becomes substantially challenging. Previous methods addressing multi-modal generalization rarely consider personalization, primarily focusing on common anatomical information. This paper aims to bridge multi-modal generalization with the concept of personalized medicine. Specifically, we propose a novel approach to derive a tractable form of the underlying personalized invariant representation $\mathbb{X}_h$ by leveraging individual-level constraints and a learnable biological prior. We demonstrate the feasibility and benefits of learning a personalized $\mathbb{X}_h$, showing that this representation is highly generalizable and transferable across various multi-modal medical tasks. Our method is rigorously validated on medical imaging modalities emphasizing both physical structure and functional information, encompassing a range of tasks that require generalization. Extensive experimental results consistently show that our approach significantly improves performance across diverse scenarios, confirming its effectiveness.

1 INTRODUCTION

Personalized medicine represents a transformative framework for 21^{st} century healthcare, tailoring medical treatments to each patient's unique characteristics (Whitcomb, 2012; Katsanis et al., 2008; Chan & Ginsburg, 2011). This approach necessitates diverse information, including clinical data such as radiological images. Three-dimensional medical images, generated through specialized techniques and radiopharmaceuticals, excel at highlighting specific anatomical features. Collectively, different medical image modalities provide a comprehensive view of a patient's structural and functional characteristics. However, this distinction between medical modalities creates significant generalization challenges in medical image analysis, especially when certain modalities may be inaccessible due to an individual's financial constraints or physical limitations, thereby complicating the effectiveness of personalized medicine.

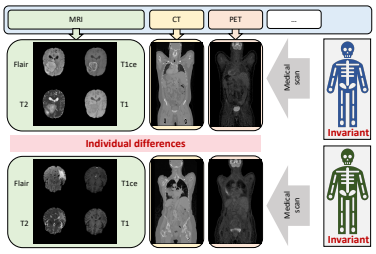

Figure 1: Diagrams of medical modalities and individual differences. The individual variations may be significant and warrant further research attention from the medical intelligence community.

As illustrated in Fig. 1, contemporary research in medical intelligence is mainly concentrated on structural modalities that depict physical anatomy. This includes Magnetic Resonance Imaging (MRI) scans (Zhao et al., 2022), which use strong magnetic fields and radiofrequency currents yielding distinct sequences, and Computed Tomography (CT) scans (Özbey et al., 2023; Zhan et al., 2024), which employ X-rays to measure its attenuation. Other studies (Yousefirizi et al., 2021) focus on the functional modalities associated with biochemistry, such as Positron Emission Tomography (PET) scans. PET scans are expensive functional imaging scans that employ radiotracers emitting

gamma rays to visualize and measure metabolic processes. The differing imaging principles result in substantial modality gaps, presenting a critical challenge for model generalization. For clarity, we categorize these modalities for generalization tasks into two types: *homogeneous generalization*, which pertains to generalizing within structural or functional modalities (e.g., T1, T2, T1ce, and Flair in MRI, as shown in Fig. 1); and *heterogeneous generalization*, which involves generalizing across both structural and functional modalities, such as CT and PET.

As each person is fundamentally different from the average of the population (Whitcomb, 2012), the concept of multi-modality generalization needs comprehensive discussion within the scope of personalization, an area scarcely addressed in previous research. An ideally well-generalized personalized medical model across modalities should (1) provide additional insights derived from all modalities when only a subset is accessible and (2) seamlessly transfer across domains while maintaining the capacity to achieve (1). While pre-training can significantly enhance downstream generalization, some recent approaches concentrate on learning common physical anatomy invariance at the class level (Jiang et al., 2023), which may overlook the individual variations. Another line of research (Tang et al., 2022; Wu et al., 2024; Jiang et al., 2023) primarily addresses the transferability of single-modal tasks and may not be suitable for multi-modal scenarios. In addition, most research on homogeneous generalization for medical tasks focuses on structural sequences of MRI or CT, employing strategies such as cross-modality transfer (Liu et al., 2023b; Kim & Park, 2024; Zhan et al., 2024) or targeting challenges like missing modality segmentation (Liu et al., 2021; Chen et al., 2023; Qiu et al., 2023a;b; Zhan et al., 2024). Heterogeneous generalization presents a greater challenge due to the disparities between structural and functional information. Despite its significance, very few efforts (Pan et al., 2023) have addressed heterogeneous generalization, mainly focusing on one-directional modality transfer (e.g., PET to CT or MRI), and rarely exploring the model's generalizability and transferability for other tasks in this context.

To address multi-modality generalization issue for personalization under both homogeneous and heterogeneous settings, we formally introduce the concept of the personalized invariant representation for multi-modal generalization, denoted as $\mathbb{X}_h$, and its constraints as outlined in Hypothesis 3.1. Furthermore, personalized invariant $\mathbb{X}_h$, which learns aggregated biological information from all possible modalities specific to the individual, is likely to enhance performance across various medical tasks for that person. Building on this hypothesis, this paper proposes a general approach aimed at enhancing the generalization of various medical imaging tasks through personalization. Specifically, our method constructs an approximation of $\mathbb{X}_h$ using the learnable biological prior knowledge $\mathbb{O}$, via decomposition, invariance, and equivariance constraints during pre-training (refer to Section 3.2). The learned approximation of $\mathbb{X}_h$ can then be utilized to enhance performance in downstream generalization tasks, irrespective of whether a domain gap exists between the pre-training data and downstream data.

Importantly, this paper demonstrates that obtaining a personalized invariant representation, $\mathbb{X}_h$, is feasible through our approach, and such invariance leads to generalization improvements across various medical tasks. To validate our methodology, we conduct experiments on modality transfer and missing modality segmentation tasks, addressing not only the homogeneous generalization of MRI but also the rarely explored heterogeneous generalization, such as generalization between PET and CT. Our findings reveal that our approach successfully captures comprehensive personalized information even when only partial modalities are available for a given individual (see Fig. 3). Moreover, extensive experiments on both homogeneous (Section 4) and heterogeneous (Section 6) generalization demonstrate that our approach can be adapted for downstream tasks and surpasses current state-of-the-art (SOTA) methods in multiple tasks, validating its superiority. We will publicly release our code, checkpoints and data upon acceptance.

2 Related work

Medical generalization tasks. Most current work focuses on homogeneous generalization, introducing tasks such as modality transfer and missing modality segmentation. The most commonly employed structural modalities — Flair, T1, T2, and T1ce of MRI — are used for brain tumor segmentation (Zhao et al., 2022), or between MRI and CT (Zhan et al., 2024) for modality transfer. Pan et al. (2023) propose an approach for heterogeneous generalization in terms of modality transfer, but only tailored for transferring PET to CT.

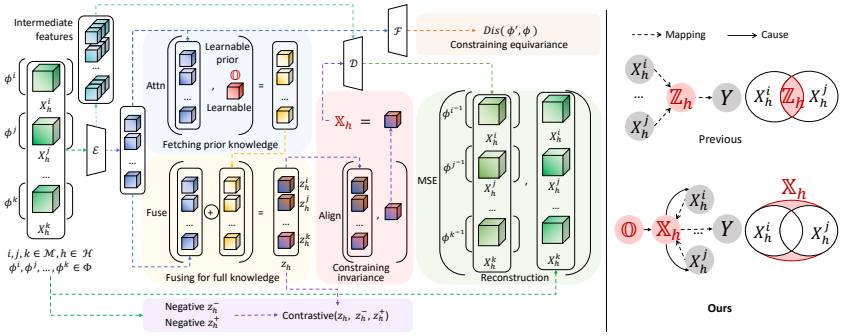

Figure 2: Left: Overall framework of learning $\mathbb{X}_h$. Right: Diagrams of differences between previous learning $\mathbb{Z}_h$ and our proposed method of learning $\mathbb{X}_h$.

Self-supervised medical pre-train models for medical generalization. Our approach aims to learn the $\mathbb{X}_h$ through pre-training, we list related medical pre-training work Tang et al. (2022); Wu et al. (2024); Chen et al. (2020b); Jiang et al. (2023) here. A notable work among them is Jiang et al. (2023), which extracts class-specific anatomical invariance. However, they only focus on a single modality. Such single-modality approaches may not be able to construct $\mathbb{X}_h$ for improving the generalization across modalities.

Generalization for medical translation. Typical modality transfer approaches are based on GAN models (Isola et al., 2017; Zhu et al., 2017; Fu et al., 2019; Park et al., 2020; Kong et al., 2021). In contrast to these GAN-based approaches, some work adopts transformer models (Liu et al., 2023b; Shi et al., 2023), while others, such as Dhariwal & Nichol (2021); Özbey et al. (2023); Kim & Park (2024); Xing et al. (2024), explore diffusion-based approaches. The methods such as MedM2G (Zhan et al., 2024) further incorporate textual information for modality transfer. Additionally, UNET-like architectures, which can also be applied to these tasks, are highlighted in (Hatamizadeh et al., 2022b;a). Most current modality transfer research focuses on improving synthesis quality. Our approach, however, demonstrates that full-modality transfer, when accompanied by specific constraints, not only enhances generation but also improves downstream generalization.

Alignment in multi-domain generalization. The issue of cross-modality generalization is similar to the problem of multi-domain generalization, which aims to extract domain invariant representations (Ganin et al., 2016; Li et al., 2018b;a; Hu et al., 2020; Tan et al., 2024). Most of these approaches focus on learning invariance across different domains, which may not fit the scope of personalization.

Generalization for medical segmentation. There are three main types of approaches to missing modality segmentation. Knowledge distillation-based approaches transfer knowledge from models with complete modality information (teachers) to models with missing modality information (students) (Chen et al., 2021; Wang et al., 2023b). (Ding et al., 2021; Zhang et al., 2022) recover missing information by leveraging the multimodal latent feature space. Domain adaptation-based methods aim to reduce the gap between models with complete and incomplete modalities by aligning their domains Wang et al. (2021). One prominent shared latent space method, MmFormer (Zhang et al., 2022), exploits intra- and inter-modality dependencies for feature fusion, which is closely related to our work. Our work reveals that our pre-train model with basic segmentation tuning exceeds these approaches.

3 Learning $\mathbb{X}_h$ for Medical Generalization

Preliminaries. In this paper, we denote the encoder as $\mathcal{E}$ and its corresponding decoder as $\mathcal{D}$. For an individual human being $h \in \mathcal{H}$, the corresponding medical images are represented as $X_h = X_h^i, X_h^j, \ldots, X_h^k$, where $i, j, \ldots, k \in \mathcal{M}$, and $\mathcal{M}$ represents the set of all possible modality combinations. We denote the intermediate features produced by $\mathcal{E}(X_h)$ and $\mathcal{E}(X_h^i)$ as x_h and x_h^i, respectively. The final layer features from the encoder are represented as z_h and z_h^i. The learned approximation of $\mathbb{X}_h$ is denoted as $\mathbb{X}_h{}'$. Additionally, we define the geometric warping function

$\phi^i \in \Phi$, where $\phi^i(X_h^i) \in \mathcal{X}$, and Φ denotes the set of all possible geometric warping functions. Finally, $I(\cdot;\cdot)$ and $P(\cdot)$ represent mutual information and probability distribution, respectively.

Before addressing the problem for both homogeneous and heterogeneous generalization, we introduce the $\mathbb{X}_h$ Hypothesis for medical imaging:

Hypothesis 3.1 ($\mathbb{X}_h$ Hypothesis)**.** Consider the set $\mathcal{M}$ of all possible modality combinations and the set Φ of all possible geometric transformations (e.g., SO(3)) transformations, for example, rotations corresponding to different poses of the person. There exists a personalized invariant representation $\mathbb{X}_h$ for an individual from the population $h \in \mathcal{H}$, which can be decomposed into modality-specific images X_h^i given a modality combination $i \in \mathcal{M}$ as a condition:

$$X_h^i = \mathbb{X}_h | i;\ i \in \mathcal{M}, h \in \mathcal{H}, \quad \textit{s.t.,} \mathbb{X}_h \perp\!\!\!\perp \mathcal{M}, \Phi. \tag{1}$$

Despite potential differences in modalities and individual variations, clinical diagnoses focus on the biological conditions of a certain patient, which remain mostly invariant during a single hospital visit. Thus, the $\mathbb{X}_h$ Hypothesis holds in most cases. Our method aims to obtain an accurate approximation of $\mathbb{X}_h$. The overall learning framework for $\mathbb{X}_h$ is illustrated on the left-hand side of Fig. 1. Data from each modality are encoded by $\mathcal{E}$, and the encoded features are used to retrieve knowledge from the learnable biological prior $\mathbb{O}$. The features and retrieved knowledge are then fused. By applying constraints of decomposition, equivariance, and invariance on the fused features, we approximate $\mathbb{X}_h$ effectively.

As illustrated in Fig. 2 right-hand side top, previous approaches (Liu et al., 2021; Chen et al., 2023; Qiu et al., 2023a;b) learn invariant representations $\mathbb{Z}_h$ across modalities through the encoder $\mathcal{E}$ for generalization: $\mathcal{E}(X_h^m) \rightarrow \mathbb{Z}_h, \mathbb{Z}_h \perp\!\!\!\perp \mathcal{M}, m \in \mathcal{M}, h \in \mathcal{H}$ during pre-training or training. Is $\mathbb{Z}_h$ a good approximation of $\mathbb{X}_h$, and does it benefit the generalization of different downstream tasks? The answer might be negative because such an approach may erase modal-specific information in $\mathbb{Z}$, making it impossible to be decomposed back into different modalities as shown in Eq. (1). Moreover, while current studies Havaei et al. (2016); Varsavsky et al. (2018); Zhang et al. (2022); Ding et al. (2021) also disentangle modality-dependent features alongside the invariant representation $\mathbb{Z}$ to enhance transferability, this strategy may compromise the generalization ability of $\mathbb{Z}$. The reason is that the transferred targets become constrained by the learned modal-dependent features, potentially limiting their broader applicability.

3.1 Using global prior $\mathbb{O}$ for better $\mathbb{X}_h$

To learn a better approximation of $\mathbb{X}_h$, we leverage a global biological prior, denoted as $\mathbb{O}$. If $\mathbb{O}$ can be learned, representations from any modality can complete themselves by retrieving the missing knowledge from $\mathbb{O}$, forming a better approximation of $\mathbb{X}_h$. Empirically, we initialize a learnable tensor as $\mathbb{O}$. As shown in Fig. 1, the representation z_h^i retrieves its missing knowledge from $\mathbb{O}$ via attention: $z_h^{i\prime} := attn(query : z_h^i, key : \mathbb{O}, value : \mathbb{O})$. The original representation and the retrieved knowledge are then fused through convolution: $\mathbb{X}_h^i := conv(z_h^{i\prime}, z_h^i)$. If the model is well-trained under the constraints of equivariance, invariance, and decomposition, the fused feature $\mathbb{X}_h^i$ becomes $\mathbb{X}_h'$, a good approximation of $\mathbb{X}_h$. The details of these constraints are discussed in Section 3.2.

3.2 Learning $\mathbb{X}_h$ by prior $\mathbb{O}$ through constraints of equivariance, invariance and decomposition

Contrastive learning. Before we introduce the constraints, we include the contrastive loss as our baseline. During the pre-training stage, we follow previous work (Chen et al., 2020b; Tang et al., 2022) and employ the contrastive learning loss. Specifically, the positive pairs are constructed as augmented samples from the same sub-volume, while the negative pairs are the views from different sub-volumes. Similar to (Tang et al., 2022), the contrastive coding is obtained by attaching a linear layer to the z_h, z_h^+, and z_h^-. Hence, the contrastive loss is then defined as:

$$\mathcal{L}_{contrast} = -\log \exp\left(\text{sim}\left(z_h, z_h^+\right)/t\right) / \exp\left(\text{sim}\left(z_h, z_h^-\right)/t\right), \tag{2}$$

where t is the measurement of the normalized temperature scale and $sim(\cdot,\cdot)$ denotes the dot product between normalized embeddings as the similarity.

As discussed in Section 3.1, the $\mathbb{X}_h$ can be obtained through a model trained under the constraints of equivariance, invariance, and decomposition. The following part presents details of those constraints according to the $\mathbb{X}_h$ hypothesis.

Invariance constraint. We constrain the invariance for $\mathbb{X}_h$ where $\mathbb{X}_h \perp\!\!\!\perp \mathcal{M}, \Phi$ through alignment. The z_h^i firstly uses attention to fetch the knowledge from the prior: $z_h^{i\prime} = attn(z_h^i, \mathbb{O})$ and then they are concatenated and fused through convolution $\mathbb{X}_h^{i\prime} = conv(z_h^i \oplus z_h^{i\prime})$. Despite the different modality combinations and geometric transformations, $\mathbb{X}_h$ should be invariant for the person:

$$\mathcal{L}_{inv} = \sum ||\mathbb{X}_h^{i\prime}, \mathbb{X}_h{}'||^2, \quad i \in \mathcal{M}. \tag{3}$$

While it is well aligned, $\mathbb{X}_h^{i\prime} = \mathbb{X}_h^{j\prime} = ... = \mathbb{X}_h{}'$ where $j \in \mathcal{M}$. Empirically, we use $\mathbb{X}_h{}' \triangleq mean(\mathbb{X}_h^{i\prime}, \mathbb{X}_h^{j\prime}, ...)$ and $mean(\cdot)$ refers the averaging of the input sequence.

Equivariance constraint. To learn better $\mathbb{O}$ and $\mathbb{X}_h{}'$ as the personalized invariant representation, we constrain the geometric equivariance and representation invariance. Consider the sample space of all modalities $X_h^i \in \mathcal{X}, i \in \mathcal{M}$, the geometric equivariance constraint forces that the geometry of the generated medical image is equivariant to ϕ^i, which can be constrained by the MSE loss in Eq. (6). Furthermore, such equivariance demands that $\phi(x_h^i)$ and z_h^i contain the information of the geometric transformation ϕ^i, inferring that it is able to extract the ϕ^i from $\phi^i(x_h^i)$ and z_h^i. Therefre, if ϕ^i can be extracted from the last-layer output z_h^i, it can also be extracted from the $\phi^i(x_h^i)$ from the previous layers:

$$\min_{\mathcal{D},\mathcal{F}} Dis(\phi^i, \mathcal{F}(z_h^i)), \tag{4}$$

where $\mathcal{F} : \mathcal{F}(z_h^i) \rightarrow \phi^{i\prime}$ extracts the geometric transformation and $Dis(\dot{;})$ denotes the distance measurement between $\phi^{i\prime}$ and ϕ^i. Empirically, following (Tang et al., 2022), we also adopt rotation as the geometric transformation, predicting the angle categories of input sub-volume is rotated. Under this case, Φ_R is defined as rotations at $[0, 90, 180, 270]$ degrees along the z-axis, and $\phi_r^i \in \Phi_R$ is the ground truth rotation categories. $\mathcal{F}z_h^i$ produces the softmax probabilities of rotation categories. The loss is in the form of:

$$\mathcal{L}_{equ} = -\sum_{r=1}^{|\Phi_R|} \phi_r^i \log \mathcal{F}(z_h^i). \tag{5}$$

Decomposition constraint. As shown in Eq. (1) of $\mathbb{X}_h$ Hypothesis, the $\mathbb{X}_h{}'$ need to be able to be decomposed as different modalities, which refers: $\min_{\mathcal{E},\mathcal{D},\mathbb{O}} I(P(\mathbb{X}_h{}'|i); P(X_h^i))$. An intuitive approach is reconstructing all possible modalities by using $\mathbb{X}_h$, whose objective can be formed as:

$$\mathcal{L}_{decom} = \sum_1^{|\mathcal{M}|} \left\| \phi^{i^{-1}} \left(\mathcal{D} \left(\mathbb{X}_h{}' | \phi^i(x_h^i) \right) \right), X_h \right\|^2, \quad i \in \mathcal{M}, \tag{6}$$

where $\phi^i(x_h^i)$ represents intermediate representations produced during $\mathcal{E}(\phi^i(X_h^i))$ and X_h denotes all possible modalities. Intuitively, $\phi^i(x_h^i)$ from earlier layers of the encoder constrains modality information thus $\mathcal{D}(\mathbb{X}_h{}'|\phi^i(x_h^i)) \triangleq \mathcal{D}(\mathbb{X}_h{}'|i)$. Specifically, the generated medical image is transformed back by using the inverse of ϕ^i to align with the inputs.

Final loss for learning $\mathbb{X}_h$. The final loss for pre-training is the combination of above losses:

$$\mathcal{L}_{pre} = \mathcal{L}_{contrast} + \mathcal{L}_{decom} + \mathcal{L}_{equ} + \mathcal{L}_{inv}, \tag{7}$$

where the weight of each loss is omitted here.

3.2.1 THE CONNECTION BETWEEN THE CONSTRAINTS AND $\mathbb{O}$

It is important to note that the above constraints are closely interconnected, as they align with Eq. (1). After obtaining additional knowledge from $\mathbb{O}$, the invariance constraint ensures that the representations from each modality for a given individual are the same, such that $\mathbb{X}_h^{i\prime}$ and $\mathbb{X}_h{}'$ can be considered equivalent. Combined with the decomposition constraint, which enforces that $\mathbb{X}_h{}'$ is shared for the generation of all possible modalities, $\mathbb{X}_h{}'$ is thus able to generalize across modalities.

Additionally, the equivariance and decomposition constraints implicitly maintain SO(3)-equivariance by satisfying the relation $\mathcal{D} \circ \mathcal{E}(\phi^i(X_h^i); \theta) = \phi^i(\mathcal{D} \circ \mathcal{E}(X_h^i; \theta))$[1], where θ represents

[1]The SO(3) transformations are left-multiplication; they are expressed here in a simplified form, using $\phi^i(\cdot)$.

Table 1: **Modality transfer results of MRI** on BRATS23: Comparison between previous methods and our method. The best results are highlighted in blue. Results denoted with * are gained from Kim & Park (2024), while the results denoted with † are gathered from Xing et al. (2024).

	Task	T1→T2			T2 → Flair			T1 → T1ce	
	Method	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	SSIM↑
2D	Pix2Pix (Isola et al., 2017)	24.624*	0.109*	0.874*	24.82†	0.0250†	0.846†	27.05†	0.858†
	CycleGAN (Zhu et al., 2017)	23.535*	0.155*	0.837*	23.418*	0.164*	0.825*	30.13†	0.906†
	NICEGAN (Chen et al., 2020a)	23.721*	0.148*	0.840*	23.643*	0.148*	0.829*	-	-
	GcGAN (Fu et al., 2019)	-	-	-	29.98†	-	0.917†	25.98†	0.872†
	CUT (Park et al., 2020)	-	-	-	23.54†	-	0.819†	26.27†	0.846†
	RegGAN (Kong et al., 2021)	24.884*	0.094*	0.881*	24.576*	0.112*	0.852*	31.36†	0.930†
	ResViT (Dalmaz et al., 2022)	25.578*	0.088*	0.895*	24.825*	0.108*	0.861*	31.46†	0.932†
	Diffusion (Dhariwal & Nichol, 2021)	-	-	-	31.98†	-	0.930†	29.22†	0.921†
	MD-Diff (Xing et al., 2024)	-	-	-	30.76†	-	0.934†	33.08†	0.948†
3D	Pix2Pix	23.740*	0.138*	0.835*	23.508*	0.152*	0.822*	-	-
	CycleGAN	25.181*	0.097*	0.887*	24.602*	0.113*	0.854*	-	-
	EaGAN (Yu et al., 2019)	24.884*	0.094*	0.881*	24.576*	0.112*	0.852*	-	-
	MS-SPADE (Kim & Park, 2024)	25.818*	0.079*	0.904*	25.074*	0.098*	0.867*	26.119*	0.912*
	Ours	**30.756**	**0.065**	**0.944**	**32.224**	**0.046**	**0.941**	**34.547**	**0.955**

the model parameters after training. This ensures that geometric transformations are preserved in the latent features z_h^i, such that $\mathcal{F}(z_h^i) = \phi^i$. The invariance constraint then requires that z_h^i with geometric transformations can retrieve features from $\mathbb{O}$ to form $\mathbb{X}_h^{i\,\prime}$, which remains invariant to any geometric transformation. This implicitly constrains $\mathbb{O}$ to contain comprehensive biological information, including other potential geometric transformations, thereby improving the $\mathbb{X}_h'$ through $\mathbb{O}$ in approximating $\mathbb{X}_h$ and enhancing the robustness of $\mathbb{X}_h{}'$.

3.3 Applying $\mathbb{X}_h$ for different modalities and tasks

After pre-training with the loss function $\mathcal{L}_{pre}$, the model is then utilized for downstream tasks such as segmentation or generation. We denote the commonly used loss functions for these tasks, such as dice loss, cross-entropy loss, or mean squared error (MSE) loss, as $\mathcal{L}_{ori}$, where paired data and labels $(X, Y) \in (\mathcal{X}, \mathcal{Y})$ are provided. In addition to $\mathcal{L}_{ori}$, we incorporate the invariance loss, denoted as $\mathcal{L}inv$, as part of the fine-tuning process for downstream tasks:

$$\mathcal{L}_{down} = \mathcal{L}_{ori} + \mathcal{L}_{inv}. \tag{8}$$

Empirically, we adopt the SwinUNETR architecture (Hatamizadeh et al., 2022a) as the backbone of the encoder $\mathcal{E}$, and implement the proposed components. The model is trained with $\mathcal{L}_{pre}$ during the pre-training phase, users have the option to either use the standard SwinUNETR by loading only our pre-trained encoder weights or to employ our proposed model structure with all pre-trained weights for downstream tasks. Notably, all modalities for a given individual, $X_h = X_h^i, X_h^j, ..., X_h^k, i, j, ..., k \in \mathcal{M}$, share the same encoder, with the encoder's channel size set to match the number of modality types. The input volume size for all experiments is fixed at $96 \times 96 \times 96$. Further empirical details on how $\mathbb{X}_h$ is leveraged for homogeneous and heterogeneous generalization are provided in Sections 4 and 6.

4 Homogeneous generalization: structural modalities of MRI

This section demonstrates that our approach enhances the homogeneous generalization across structural modalities in MRI. To validate that our method captures personalized information, especially anatomical structure features, during the pre-training stage, we first apply it to modality transfer tasks. Next, we adapt the pre-trained model for the downstream missing modality segmentation task. Experimental results indicate that our approach outperforms state-of-the-art (SOTA) methods in both tasks, thereby supporting the $\mathbb{X}_h$ Hypothesis and confirming the effectiveness of our method.

4.1 Pre-training for modality transfer

Modality transfer tasks focus on converting medical images from multiple modalities to other modalities. We test our approach on the structural modalities of MRI. This task aligns seamlessly with our pre-training objective, and the quality of the generated modalities serves as a validation of the

Table 2: **Modality transfer results of MRI** on BRATS23: Comparison between the previous method and ours for transfer between all four modalities. The averaged results of metrics for all validation samples are listed. The best results are highlighted in blue. Please refer to Appendix 9 for the standard derivations according to each result. Please refer to Appendix Table 9 for standard derivations of the results.

Source	Target	T1			T1ce			T2			Flair		
		PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑
T1	SwinUNETR	32.815	0.092	0.941	31.655	0.202	0.912	24.650	0.361	0.857	27.593	0.202	0.883
	MS-SPADE	29.001	0.055	0.942	26.119	0.078	0.912	25.818	0.103	0.904	24.842	0.113	0.859
	Ours	**43.472**	**0.003**	**0.996**	**34.547**	**0.045**	**0.955**	**30.756**	**0.065**	**0.944**	**31.693**	**0.049**	**0.937**
T1ce	SwinUNETR	32.456	0.100	0.929	33.001	0.156	0.926	25.125	0.366	0.859	27.699	0.211	0.882
	MS-SPADE	26.228	0.076	0.922	28.759	0.060	0.937	25.990	0.092	0.907	25.204	0.092	0.881
	Ours	**34.077**	**0.020**	**0.962**	**46.663**	**0.003**	**0.996**	**30.775**	**0.063**	**0.942**	**32.224**	**0.046**	**0.941**
T2	SwinUNETR	30.102	0.171	0.896	30.354	0.283	0.883	26.831	0.268	0.887	27.234	0.242	0.872
	MS-SPADE	25.422	0.085	0.908	25.234	0.087	0.895	29.230	0.048	0.942	25.074	0.098	0.867
	Ours	**32.646**	**0.028**	**0.955**	**33.857**	**0.051**	**0.949**	**43.653**	**0.006**	**0.991**	**32.224**	**0.046**	**0.941**
Flair	SwinUNETR	31.371	0.135	0.916	31.285	0.240	0.905	25.579	0.338	0.867	29.092	0.148	0.923
	MS-SPADE	25.186	0.090	0.905	25.899	0.094	0.906	26.146	0.086	0.913	28.608	0.058	0.938
	Ours	**32.752**	**0.026**	**0.951**	**33.471**	**0.055**	**0.944**	**30.571**	**0.068**	**0.940**	**43.624**	**0.004**	**0.995**

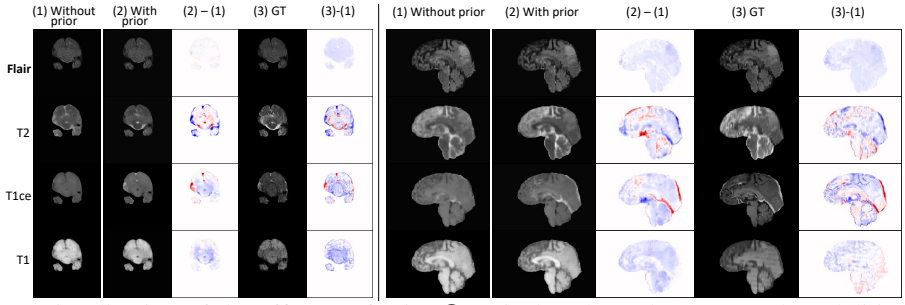

Figure 3: Visualization of the efficacy of prior $\mathbb{O}$. Displayed are the generated modalities on the input Flair modality of a testing sample on the BTATS21 dataset. Columns show: the generated images of the model (1) without prior $\mathbb{O}$ and (2) with prior $\mathbb{O}$ are aligned with (3) the GT images. Typically, the differences between without and with prior $\mathbb{O}$ (the (2)-(1) column) are visualized to compare with the differences between without $\mathbb{O}$ and GT (the (3)-(1) column). Red and blue refer to the positive (accomplishment) and negative (refinement) values of the differences, respectively.

anatomical knowledge captured by our pre-trained model. Importantly, our approach can generate all modalities without knowing the exact modalities where the input form, as the learned representation $\mathbb{X}_h{}'$ encompasses comprehensive information across all possible modalities.

Experimental settings. Following previous methods (Kim & Park, 2024), we utilize the multimodal brain tumor segmentation challenge 2023 (BRATS23) dataset (Baid et al., 2021; Menze et al., 2014; Bakas et al., 2017b;a). BRATS23 includes four structural MRI modalities (T1, T1ce, T2, and FLAIR) for each individual. Our model is tested on the BRATS23 validation set, which contains these four modalities for 219 individuals. We evaluate the quality of synthesis using peak signal-to-noise ratio (PSNR), normalized mean squared error (NMSE), and structural similarity index (SSIM) (Yi et al., 2019). To provide comprehensive results, we separately compare the translation results for T1 $\rightarrow$ T2 and T2 $\rightarrow$ FLAIR, as some previous methods are only capable of single-modality transfer. These include both 2D and 3D generation methods, as shown in Table 1. Additionally, we employed SwinUNTER for multi-modality translation comparisons. All evaluations were performed on 3D volumes; for the 2D methods, synthesized target images were stacked to form a 3D volume for comparison. Please refer to model training details to Appendix B.

Results. Table 1 exhibits the transfer results of T1 $\rightarrow$ T2 and T2 $\rightarrow$ Flair. Our approach significantly surpasses previous 2D and 3D generation methods, including single- and multi-modality translation methods. Specifically, our approach exceeds current SOTA diffusion-based methods, such as 2D-based MD-Diff and 3D-based MS-SPADE. In terms of multi-modality translation, as Table 2 shows, our approach performs better than MS-SPADE and SwinUNETR across all metrics under all settings. Moreover, for all settings, it can be seen that our method significantly improves the SSIM, indicating a better anatomy structure obtained by our approach. These results indicate that the $\mathbb{X}_h$ Hypothesis is plausible for homogeneous generalization, and our personalized approach is able to obtain its approximation.

Analysis of $\mathbb{O}$. We show that using $\mathbb{O}$ for $\mathbb{X}_h$ mainly accomplishes the personalized knowledge of each sequence from MRI modalities. Those modalities are mainly focused on the physical anatomy.

Table 3: **Missing modality segmentation results of MRI** on BRATS18: Num denotes the number of missing modalities for different settings. We report mean and standard deviations of DICE results of all experimental combinations under the same Num. The best results are highlighted in blue. Please refer to Appendix Table 10 for detailed results of each setting.

Method		All settings		Full Modality	Missing Num =1		Missing Num =2		Missing Num =3	
		Mean	Std.	-	Mean	Std.	Mean	Std.	Mean	Std.
Tumor core	RFNET (Ding et al., 2021)	76.08	6.99	83.40	80.63	4.53	76.57	7.15	68.95	6.07
	mmFormer Zhang et al. (2022)	76.43	5.83	82.22	79.78	4.33	76.55	6.03	71.45	6.80
	SPA (Wang et al., 2023a)	74.80	6.95	82.23	78.99	5.38	75.01	8.31	68.44	8.88
	M3AE (Liu et al., 2023a)	72.67	7.43	80.29	77.61	4.59	73.37	7.96	64.79	9.85
	M2F (Shi et al., 2023)	73.69	6.83	80.34	77.48	5.19	74.17	6.88	67.51	6.63
	Ours	**79.78**	**4.55**	**86.72**	**83.64**	**2.29**	**79.56**	**3.28**	**74.51**	**3.87**
Enhancing tumor	RFNET	59.31	15.10	73.65	**66.91**	12.90	59.17	14.50	48.35	13.86
	mmFormer	62.14	18.60	**79.91**	71.54	15.94	61.77	19.65	48.86	20.73
	SPA	58.92	17.68	73.40	68.44	17.11	58.05	19.77	47.10	19.99
	M3AE	55.98	17.45	73.79	65.09	14.54	55.53	20.37	43.09	22.03
	M2F	58.84	16.58	75.26	66.67	13.83	58.99	16.19	46.70	15.97
	Ours	**63.49**	**7.58**	70.64	64.44	**6.62**	**63.87**	**11.45**	**60.19**	**11.31**
Whole tumor	RFNET	83.92	6.14	89.27	87.25	2.94	84.95	3.81	77.70	7.46
	mmFormer	84.84	5.35	88.26	87.59	2.40	85.36	5.60	80.45	4.64
	SPA	84.52	5.48	89.03	87.81	**1.25**	85.26	4.69	78.98	7.32
	M3AE	81.52	6.71	86.82	85.64	1.36	82.43	6.08	74.74	8.69
	M2F	83.88	5.79	88.72	87.30	1.99	84.62	2.57	78.13	7.81
	Ours	**87.63**	**3.25**	**91.19**	**89.49**	1.82	**87.45**	**1.03**	**85.17**	**3.93**

For the Flair modality in MRI, which mainly highlights the lesion but suppresses structures like bones, Fig. 3 shows that without $\mathbb{O}$, the main difference between the generated images and ground truth (GT) images is the personalized structure. Prior $\mathbb{O}$ for $\mathbb{X}_h$ accomplishes and refines the personal level anatomical information, mitigating the gap between them with the GT, so it can be better transferred to other structural focusing modalities.

4.2 Tuning for missing modality segmentation

Experimental settings. To validate the generalization ability of the pre-trained model, we fine-tune the model obtained from Section 4.1 on the BRATS18 (Menze et al., 2014) from the Multimodal Brain Tumor Segmentation Challenge. Similar to BRATS23, BRATS18 also consists of the same four structural modalities. We employ the Dice similarity coefficient (DICE) as the metric for evaluation. For a fair comparison, we follow data splits of Shi et al. (2023) and reproduce the results of previous methods () on these splits by using their released code and following their original settings [2]. Additional experimental details can be seen in Appendix B.

Results. Table 3 presents the segmentation results of our approach compared to previous methods. We also compute the standard deviation of DICE scores under various missing modality settings, which highlights the robustness of our model. Notably, our approach outperforms previous methods in most missing modality scenarios, particularly when the number of missing modalities is large. Our method shows significant improvements in enhancing tumor segmentation, especially when Missing Num = 3. Moreover, the reduced standard deviation of DICE scores under different missing modality settings indicates that our personalized approach consistently delivers superior segmentation results. This performance improvement stems from the enhanced generalization of our model, which is rooted in the learned $\mathbb{X}_h{}'$.

5 Heterogeneous generalization: PET and CT modalities

Given the differing imaging principles, the modality gap in heterogeneous generalization may be more pronounced than that in homogeneous generalization, making the former tasks more challenging. In this section, we evaluate our approach for heterogeneous generalization.

5.1 Pre-training for modality transfer

Experimental settings. We utilize the AutoPET-II dataset Gatidis et al. (2022) from the Automated Lesion Segmentation in PET/CT challenge for pre-training. The AutoPET-II dataset includes AC-

[2]Though we tried our best, it can be noticed some reproduced results are lowered than their reported results in their original paper. It should be clarified that **our results also exceed those reported results**. However, for a comprehensive study, we mainly report our reproduced results.

Table 4: **Ablation study - Modality transfer results of PET and CT** on AutoPET-II: Ablation results of models trained under different combinations of constraints. The best and second results are highlighted in blue and cyan, respectively.

	SSIM↑						PSNR↑					
+ Contrastive + Decomposition	•	•	•	•	•	•	•	•	•	•	•	•
+ Equivariance	•		•	•		•	•		•	•		•
+ Invariance		•	•		•	•		•	•		•	•
+ 𝕆				•	•	•				•	•	•
PET → PET	0.9903	0.9835	0.9955	0.9931	0.9957	**0.9969**	44.8811	42.2223	46.5603	45.5829	**47.4198**	**49.5473**
CT → CT	0.9739	0.9475	0.9419	0.9437	0.9664	**0.9780**	37.2309	32.0777	31.2692	33.1866	35.4194	**37.0989**
PET → CT	0.9161	0.9148	0.9215	0.9070	0.9121	**0.9282**	28.1046	29.3181	**29.6694**	26.8885	27.2708	**30.1548**
CT → PET	**0.9884**	0.9824	0.9851	0.9834	0.9842	0.9883	39.8490	39.0795	39.1718	39.4528	39.4348	**41.5840**
Avg.	0.9672	0.9571	0.9610	0.9568	0.9646	**0.9728**	37.5164	35.6744	36.6677	36.2777	37.3862	**39.5963**

Table 5: **Segmentation results of PET and CT** on AutoPET-II: Comparison between the previous method and ours. The best results are highlighted in blue.

Method	Dice↑	Dice-↑	TPR↑	TNR↑	FNR↓	FPR↓
nnUnet (Isensee et al., 2021)	33.10	-	-	-	-	-
UNETR (Hatamizadeh et al., 2022b)	10.81	23.14	80.65	3.64	19.35	96.36
SwinUNETR (Hatamizadeh et al., 2022a) without pre-train	43.45	62.60	**90.32**	62.73	**9.68**	37.27
SwinUNETR with its pre-train (Tang et al., 2022)	44.06	57.79	89.25	73.64	10.75	26.36
Ours	**48.20**	**61.16**	88.17	**77.27**	11.83	**22.72**

PET and CT pairs, where the PET scans adopt FDG tracers, and their attenuation is corrected using the corresponding CT scans. Specifically, we divide the AutoPET-II dataset into training and testing sets. Similar to our approach for heterogeneous generalization, we adopt the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) as evaluation metrics. In this section, we present the results of models that employ different combinations of the constraints and the set 𝕆. We use both contrastive loss and the decomposition constraint as our baseline. Please refer to training details in Appendix B.

Results. As shown in Table 4 and generated examples in Fig. 4, incorporating 𝕆 with different combinations of constraints improves generation quality across most metrics. Specifically, using the constraints without 𝕆 does not guarantee improvements, as discussed in Section 3.2.1. Ultimately, employing all constraints along with 𝕆 yields the best average results across all translations, validating that our approach performs well in heterogeneous generalization settings. These results indicate that our method under the scope of personalization bridges the gap between structural and functional modalities. We validate the transferability of all these pre-train models in Appendix A.1, where additionally analysis are provided.

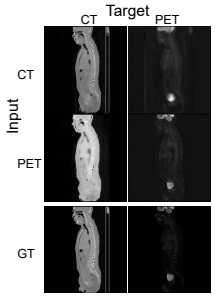

Figure 4: Generated samples on AUTOPET-II.

5.2 Tuning for Segmentation

Experimental settings. We utilize the AutoPET-II Gatidis et al. (2022) dataset for segmentation, evaluating performance using the DICE metric. It is important to note that we employ the same training and testing splits as in Section 5.1 to avoid data leakage. Specifically, we adhere to the settings from the official challenge; DICE is calculated in the standard manner but is set to zero for false negatives and true negatives. Additionally, we introduce DICE- to include the mean across all samples, along with true positive rate (TPR), true negative rate (TNR), false negative rate (FNR), and false positive rate (FPR) for the missing modality segmentation evaluation. Our method is compared against nnUNET (Isensee et al., 2021), UNETR (Hatamizadeh et al., 2022b), and SwinUNETR (Hatamizadeh et al., 2022a), which are trained directly on the dataset without pre-training. Notably, we also compare our approach with SwinUNETR using its pre-training strategy (Tang et al., 2022). Please refer to training details in Appendix B.

Results. The full modality segmentation results are exhibited in Table 5. The results indicate that with proper model architecture, such as SwinUNETR, using both two modalities usually outperforms solely using PET. It can be observed that models using our pre-train improve the results across all metrics. Typically, SwinUNETR using our pre-train significantly exceeds it without our pre-trained model, indicating the personalized invariant learned by our pre-train generalizes to the downstream well and can boost the downstream tasks. Moreover, using our proposed components

Table 6: **Modality transfer results of NAC-PET to AC-PET and CT** that tuned on HNSCC and evaluated on HNSCC validation set and NSCLC: Comparison between the previous method and ours for transfer between different modalities. The best results are highlighted in blue.

	SSIM↑			PSNR↑		
HNSCC validation	NAC→CT	NAC →AC	Avg.	NAC→CT	NAC→AC	Avg.
UNETR	0.4899	0.8998	0.6949	21.7330	42.8557	32.2944
SwinUNETR	0.5853	0.9265	0.7559	23.5628	42.5495	33.0561
Ours	**0.6939**	**0.9516**	**0.8227**	**25.8498**	**46.4658**	**36.1578**
	SSIM↑			PSNR↑		
NSCLC	NAC→CT	NAC →AC	Avg.	NAC→CT	NAC→AC	Avg.
UNETR	0.4476	0.8703	0.6590	20.6182	40.8570	30.7376
SwinUNETR	0.4476	0.8705	0.6591	22.5086	41.3272	31.9179
Ours	**0.4744**	**0.8853**	**0.6798**	**22.7791**	**42.7687**	**32.7739**

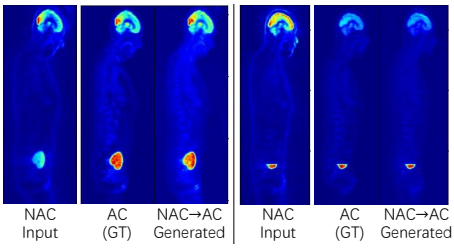

Figure 5: **Modality transfer results of NAC-PET to AC-PET:** Generated examples on the NSCLC dataset for NAC → AC across individuals.

with the pre-train leads to the best DICE and DICE-. This validates that using the prior further emphasizes the personalized invariant, which yields the most segmentation improvements.

6 Special Case: A More Complex Scenario

We introduce a more complex scenario, in which the pre-trained model for heterogeneous generalization settings is tuned downstream that span both heterogeneous and homogeneous generalization.

Experimental settings. The pre-train model we adopted is from that trained on AC-PET and CT. Specifically, we tune the model by using the Head and Neck Squamous Cell Carcinoma (HNCSS) dataset (Grossberg et al., 2020) as the training set and the Non-Small Cell Lung Cancer (NSCLC) dataset as the testing set. Both datasets are sourced from The Cancer Imaging Archive (TCIA) (Clark et al., 2013), and they contain paired non-attenuation-corrected PET (NAC-PET), attenuation-corrected PET (AC-PET), and CT scans. The model is pre-trained for heterogeneous generalization between AC-PET and CT. It is tuned for both homogeneous generalization between AC-PET and NAC-PET and heterogeneous generalization between NAC-PET and CT. Similar to the previous translation experiments, we use SSIM and PSNR as evaluation metrics. Performance in this scenario further validates the model's generalization capabilities. Note here the training and testing data in the downstream task come from different domains. See training details in Appendix B.

Results. Table 6 presents the results on the HNSCC dataset, while Fig. 5 displays generated sample images for homogeneous generalization. Our approach achieves superior results across both heterogeneous and homogeneous generalizations. For heterogeneous generalization, our method consistently improves SSIM for NAC-PET to CT, indicating that the learned $\mathbb{X}_h{}'$ successfully captures and emphasizes anatomical structures in the generated images, as indicated by improved SSIM. Moreover, though the model is pre-trained between AC-PET and CT, the improvements are also consistent for NAC-PET and AC-PET. These findings confirm that our personalized approach is effective for a complex real-world scenario, demonstrating the transferability and generalizability of the pre-trained model to downstream tasks under various scenarios.

7 Conclusion

This paper proposes a universal approach to tackle multi-modality generalization by approximating personalized invariant representation $\mathbb{X}_h$ through invariance, equivariance, and decomposition constraints with a learnable biological prior. We specifically unveil that learning $\mathbb{X}_h$ is feasible, and it would significantly benefit the generalization in medical tasks.

Limitations, challenges, and future work. To enhance the validation of our approach, we adhere to commonly used settings during the tuning stage. Exploring alternative strategies, such as knowledge distillation, could further improve downstream performance. Our approach requires datasets where all modalities are instance-level matched, which can be a stringent condition and may be unattainable for certain modalities. Future research should explore methods to achieve personalized invariance without relying on instance-level matched datasets. Additionally, we advocate for the availability of more open-source multi-modal medical datasets, particularly for functional modalities, as these are not widely accessible to researchers.

REFERENCES

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. *The cancer imaging archive*, 286, 2017a.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4 (1):1–13, 2017b.

Isaac S Chan and Geoffrey S Ginsburg. Personalized medicine: progress and promise. *Annual review of genomics and human genetics*, 12(1):217–244, 2011.

Cheng Chen, Qi Dou, Yueming Jin, Quande Liu, and Pheng Ann Heng. Learning with privileged multimodal knowledge for unimodal segmentation. *IEEE transactions on medical imaging*, 41 (3):621–632, 2021.

Delin Chen, Yansheng Qiu, and Zheng Wang. Query re-training for modality-gnostic incomplete multi-modal brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 135–146. Springer, 2023.

Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8168–8177, 2020a.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26:1045–1057, 2013.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yuhang Ding, Xin Yu, and Yi Yang. Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3975–3984, 2021.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2427–2436, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

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):601, 2022.

Aaron Grossberg, Hesham Elhalawani, Abdallah Mohamed, Sam Mulder, Bowman Williams, Aubrey L White, James Zafereo, Andrew J Wong, Joel E Berends, Shady AboHashem, et al. Md anderson cancer center head and neck quantitative imaging working group. *HNSCC [Dataset]. The Cancer Imaging Archive*, 2020.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022a.

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 *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 574–584, 2022b.

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*, pp. 469–477. Springer, 2016.

Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*, pp. 292–302. PMLR, 2020.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, and Minfeng Xu. Anatomical invariance modeling and semantic alignment for self-supervised learning in 3d medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15859–15869, 2023.

SH Katsanis, Gail Javitt, and Kathy Hudson. A case study of personalized medicine, 2008.

Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7604–7613, 2024.

Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34: 1964–1978, 2021.

Ya Li, Mingming Gong, Xinmei Tian, Tongliang Liu, and Dacheng Tao. Domain generalization via conditional invariant representations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 624–639, 2018b.

Hong Liu, Dong Wei, Donghuan Lu, Jinghan Sun, Liansheng Wang, and Yefeng Zheng. M3ae: multimodal representation learning for brain tumor segmentation with missing modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1657–1665, 2023a.

Jiang Liu, Srivathsa Pasumarthi, Ben Duffy, Enhao Gong, Keshav Datta, and Greg Zaharchuk. One model to synthesize them all: Multi-contrast multi-scale transformer for missing data imputation. *IEEE Transactions on Medical Imaging*, 42(9):2577–2591, 2023b.

Yanbei Liu, Lianxi Fan, Changqing Zhang, Tao Zhou, Zhitao Xiao, Lei Geng, and Dinggang Shen. Incomplete multi-modal representation learning for alzheimer's disease diagnosis. *Medical Image Analysis*, 69:101953, 2021.

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.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

Yongsheng Pan, Feihong Liu, Caiwen Jiang, Jiawei Huang, Yong Xia, and Dinggang Shen. Revealing anatomical structures in pet to generate ct for attenuation correction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 24–33. Springer, 2023.

Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345. Springer, 2020.

Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, and Zheng Wang. Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21317–21326, 2023a.

Yansheng Qiu, Ziyuan Zhao, Hongdou Yao, Delin Chen, and Zheng Wang. Modal-aware visual prompting for incomplete multi-modal brain tumor segmentation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3228–3239, 2023b.

Junjie Shi, Li Yu, Qimin Cheng, Xin Yang, Kwang-Ting Cheng, and Zengqiang Yan. M^2ftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2023.

Zhaorui Tan, Xi Yang, and Kaizhu Huang. Rethinking multi-domain generalization with a general learning objective. *arXiv preprint arXiv:2402.18853*, 2024.

Yucheng Tang, Dong Yang, Wenqi Li, Holger R Roth, Bennett Landman, Daguang Xu, Vishwesh Nath, and Ali Hatamizadeh. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20730–20740, 2022.

Thomas Varsavsky, Zach Eaton-Rosen, Carole H Sudre, Parashkev Nachev, and M Jorge Cardoso. Pimms: permutation invariant multi-modal segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, pp. 201–209. Springer, 2018.

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*, pp. 15878–15887, 2023a.

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)*, pp. 1–5. IEEE, 2023b.

Yixin Wang, Yang Zhang, Yang Liu, Zihao Lin, Jiang Tian, Cheng Zhong, Zhongchao Shi, Jianping Fan, and Zhiqiang He. Acn: adversarial co-training network for brain tumor segmentation with missing modalities. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24*, pp. 410–420. Springer, 2021.

David C Whitcomb. What is personalized medicine and what should it replace? *Nature reviews Gastroenterology & hepatology*, 9(7):418–424, 2012.

Linshan Wu, Jiaxin Zhuang, and Hao Chen. Voco: A simple-yet-effective volume contrastive learning framework for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22873–22882, 2024.

Zhaohu Xing, Sicheng Yang, Sixiang Chen, Tian Ye, Yijun Yang, Jing Qin, and Lei Zhu. Cross-conditioned diffusion model for medical image to image translation. *arXiv preprint arXiv:2409.08500*, 2024.

Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

Fereshteh Yousefirizi, Abhinav K Jha, Julia Brosch-Lenz, Babak Saboury, and Arman Rahmim. Toward high-throughput artificial intelligence-based segmentation in oncological pet imaging. *PET clinics*, 16(4):577–596, 2021.

Biting Yu, Luping Zhou, Lei Wang, Yinghuan Shi, Jurgen Fripp, and Pierrick Bourgeat. Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis. *IEEE transactions on medical imaging*, 38(7):1750–1762, 2019.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11502–11512, 2024.

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

Zechen Zhao, Heran Yang, and Jian Sun. Modality-adaptive feature interaction for brain tumor segmentation with missing modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 183–192. Springer, 2022.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A SOCIAL IMPACT

This work presents an approach to tackle multi-modality generalization through personalization. We hope our work can encourage the community to work towards practical, personalized medical models with border generalization ability.

A.1 DOWNSTREAM SEGMENTATION ABLATION STUDY

Table 7: **Ablation study - Segmentation results of using different pre-train models** on AutoPET-II: Comparison between the pre-train models with different settings and ours. The best results are highlighted in blue and cyan.

ID	Pretrian	**DICE**↑	DICE-↑	TPR↑	TNR↑	FNR↓	FPR↓
1	+ Contrastive + Decomposition + Equivariance	40.85	55.79	81.72	69.09	18.28	30.91
2	+ Contrastive + Decomposition + Invariance	44.34	48.63	77.42	91.82	22.58	**8.18**
3	+ Contrastive + Decomposition + Equivariance + Invariance	42.42	60.67	**89.25**	63.64	**10.75**	36.36
4	+ Contrastive + Decomposition + Equivariance + $\mathbb{O}$	46.31	55.77	83.87	82.73	16.13	17.27
5	+ Contrastive + Decomposition + Invariance + $\mathbb{O}$	44.42	57.80	88.17	74.55	11.83	25.45
6	**+ Contrastive + Decomposition + Equivariance + Invariance + $\mathbb{O}$**	**48.20**	**61.16**	**88.17**	**77.27**	**11.83**	22.72

We demonstrate the effectiveness of our proposed components and discuss the process of learning an anatomy-invariant representation. Experimental results for downstream segmentation tasks and visualizations of the pre-trained models are presented in Table 7. All experiments are conducted under consistent settings to ensure a fair comparison.

Using all constraints together with $\mathbb{O}$ yields the best results. Consistent with Section 3.2.1, the results indicate that using different constraints alone may not guarantee improvements; however, incorporating all constraints along with $\mathbb{O}$ results in the best outcomes. This validates the plausibility of the $\mathbb{X}_h$ Hypothesis and demonstrates that achieving good approximation of it significantly enhances generalization.

Using prior $\mathbb{O}$ with decomposition constraint improves the model performance for different settings. Despite different settings, additionally using $\mathbb{O}$ with decomposition improves the downstream model performance. Combined with the improvements from modality transfer results in Table 4, it suggests that $\mathbb{O}$ helps with better obtaining anatomical structure.

The invariance and equivariance constraints can not be applied to the same feature. It needs to be highlighted that invariance and equivariance constraints can not be applied to the same features as they conflict with each other. As shown in task 3, without $\mathbb{O}$, invariance and equivariance constraints are applied to the latent feature simultaneously, leading to a significant performance drop. In comparison, apply equivariance constraint before using $\mathbb{O}$ and applying invariance constraint after using $\mathbb{O}$ yields the best results. This is because the geometrical transformation contained in z_h^i needs to be accomplished by fetching other possible geometrical transformation information form $\mathbb{O}$ and then fused to be invariant.

B EXPERIMENTAL DETAILS

The model and data loaders are built by using MONAI `https://docs.monai.io/en/stable/index.html`. Please refer to all the details of the implementation in the code. We present some key implementations below.

B.1 OVERALL TRAINING PROCEDURE

We provide a pseudo-code for our approach. The loss calculation for **Pre-training** procedure is simplified as Algorithm 1 and **Downstream tuning** as Algorithm 2. It is notable that the empirical procedure is flexible as long as the $\mathbb{O}$ is properly used to construct $\mathbb{X}'_h$ and those constraints are applied to $\mathbb{X}'_h$.

Algorithm 1: Calculate losses during one step for pre-training

Data: $X \in \mathcal{X}, epoch$
Initialize learnable $\mathbb{O}$ $\mathcal{E}(\cdot), \mathcal{D}(\cdot)$;
while $i \neq epoch$ **do**
  $X'_h \leftarrow None$;
  for $h \in \mathcal{H}$ **do**
    for $i \in \mathcal{M}$ **do**
      $\mathcal{L}_{pre} \leftarrow 0$;
      $X_h^i \sim X, \phi^i \sim \Phi$;
      $X_h^{i+}, X_h^{i-} = Augment(\phi^i(X_h^i))$;
      $(z_h^i, x_h^i), (z_h^{i-}, x_h^{i-}), (z_h^{i+}, x_h^{i+}) \leftarrow \mathcal{E}(\phi^i(X_h^i)), \mathcal{E}(X_h^{i-}), \mathcal{E}(X_h^{i+})$;
      Calculate $\mathcal{L}_{contrast}(z_h^i, z_h^{i+}, z_h^{i-})$, $\mathcal{L}_{pre} += \mathcal{L}_{contrast}$;
      $\mathcal{F}(z_h^i) \rightarrow \phi^{i'}$;
      Calculate $\mathcal{L}_{equ}(\phi^{i'}, \phi^i)$, $\mathcal{L}_{pre} += \mathcal{L}_{equ}$;
      $z_h^{i\prime} := Attn(query : z_h^i, key : \mathbb{O}, value : \mathbb{O})$;
      $X_h^{i\prime} := Conv(z_h^{i\prime}, z_h^i)$;
      if X'_h *is not None ;* /* For saving memory */
       then
        Calculate $\mathcal{L}_{inv}(X_h^{i\prime}, X'_h)$, $\mathcal{L}_{pre} += \mathcal{L}_{inv}$;
        $X'_h := (X_h + X_h^{i\prime})/2$;
      else
        $X'_h := X_h^{i\prime}$;
      end
      $X_h^{i\prime} := \mathcal{D}(X_h^{i\prime}, x_h^i)$;
      Calculate $\mathcal{L}_{decom}(\phi^{i-1}(X_h^{i\prime}), X_h)$, $\mathcal{L}_{pre} += \mathcal{L}_{decom}$;
    end
  end
end

B.2 Homogeneous generalization: structural modalities in MRI

B.2.1 Pre-training and Modality transfer.

Experimental settings. We use four A100 GPUs for training. The learning rate we used for the modality transfer is set to 0.0002, and the training epoch is set to 1000. Both the number of input and out channels is set as 4.

Training details. For the model, both the input and output channels are set to 4, corresponding to the four MRI modalities. All modalities are loaded and cropped to a size of $96 \times 96 \times 96$ simultaneously. Following (Kim & Park, 2024), we also normalize each MRI modality to have zero mean and unit variance. During training, the background is excluded for modal generation. A single modality is repeated four times to create four channels during training to obtain $\mathbb{X}_h^{i\prime}$. The training loss follows the $\mathcal{L}_{pre}$, whose calculation details during the training phase can be seen in Algorithm 1.

B.2.2 Missing modality segmentation.

We use four A100 GPUs for tuning. The learning rate we used for the modality transfer is set to 0.0002, and the training epoch is set to 1000. Both the number of input and out channels is set as 4.

Training details. Following Shi et al. (2023), we also normalize each MRI modality to zero mean and unit variance. For the fine-tuning, we employ Dice loss, the weighted cross-entropy loss that is adopted by Shi et al. (2023), and the additional $\mathcal{L}_{inv}$.

Algorithm 2: Calculate losses during one step for fine-tuning

Data: $(X, Y) \in (\mathcal{X}, \mathcal{Y}), epoch$
Load pre-trained $\mathbb{O}$ $\mathcal{E}(\cdot), \mathcal{D}(\cdot)$;
while $i \neq epoch$ **do**
  $X'_h \leftarrow None$;
  for $h \in \mathcal{H}$ **do**
    for $i \in \mathcal{M}$ **do**
      $\mathcal{L}_{down} \leftarrow 0$;
      $(X^i_h, Y_h) \sim X, Y$;
      $(z^i_h, x^i_h) \leftarrow \mathcal{E}(X^i_h)$;
      $z^{i\prime}_h := Attn(query : z^i_h, key : \mathbb{O}, value : \mathbb{O})$;
      $X^{i\prime}_h := Conv(z^{i\prime}_h, z^i_h)$;
      if X'_h *is not None ;* `/* For saving memory */`
       then
        Calculate $\mathcal{L}_{inv}(X^{i\prime}_h, X'_h)$, $\mathcal{L}_{down} += \mathcal{L}_{inv}$;
        $X'_h := (X_h + X^{i\prime}_h)/2$;
      else
        $X'_h := X^{i\prime}_h$;
      end
      $Y'_h := \mathcal{D}(X^{i\prime}_h, x^i_h)$;
      Calculate $\mathcal{L}_{ori}(Y'_h, Y_h)$, $\mathcal{L}_{down} += \mathcal{L}_{ori}$;
    end
  end
end

B.3 Heterogeneous generalization: PET and CT modalities

B.3.1 Modality transfer

All models are trained using A100 GPUs. **Training details.** All models are trained under the same situations, using the same data pre-processing transforms.

B.3.2 Downstream segmentation

Training details. All training and fine-tuning experiments use the same losses, while the approaches with our pre-train additionally use $\mathcal{L}_{inv}$ for downstream fine-tuning. Moreover, we also compare the original architecture of SwinUNETR using our pre-trained weights with fully using our architecture and our weights for fine-tuning.

B.4 Special case: Tuning from heterogeneous to homogeneous generalization with domain gap

Training details. For the fine-tuning stage, we use the decoder architecture of SwinUNETR, which is randomly initialized. The training procedure is similar to the above modality transfer experiments, with the primary difference being that the input and output channels are set to two. Additionally, we reproduced the results of UNETR and SwinUNETR for comparison, ensuring that the same loss functions were applied across models.

C More results

C.1 Modality transfer results on BRATS22:

Table 8 and Table 9 presents the generation result with standard derivations. The results of our method and SwinUNETR are produced by ourselves, while the rest of the results are gathered from Kim & Park (2024). Generated examples are presented in Figs. 6 to 8.

Table 8: **Modality transfer results of MRI** on BRATS23: Comparison between previous methods and our method. The best results are highlighted in blue.

	Task	T1→T2			T2 → Flair		
Dimension	Method	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑
2D	Pix2Pix	24.624 ± 0.962	0.109 ± 0.028	0.874 ± 0.015	24.361 ± 1.061	0.117 ± 0.021	0.846 ± 0.019
	CycleGAN	23.535 ± 1.334	0.155 ± 0.035	0.837 ± 0.028	23.418 ± 0.944	0.164 ± 0.033	0.825 ± 0.035
	NICEGAN	23.721 ± 1.136	0.148 ± 0.029	0.840 ± 0.029	23.643 ± 1.045	0.148 ± 0.022	0.829 ± 0.033
	RegGAN	24.884 ± 0.991	0.094 ± 0.024	0.881 ± 0.017	24.576 ± 1.073	0.112 ± 0.022	0.852 ± 0.028
	ResViT	25.578 ± 0.812	0.088 ± 0.021	0.895 ± 0.018	24.825 ± 1.030	0.108 ± 0.018	0.861 ± 0.021
3D	CycleGAN	25.181 ± 0.861	0.097 ± 0.031	0.887 ± 0.012	24.602 ± 1.181	0.113 ± 0.021	0.854 ± 0.018
	Pix2Pix	23.740 ± 1.198	0.138 ± 0.032	0.835 ± 0.019	23.508 ± 1.301	0.152 ± 0.039	0.822 ± 0.024
	EaGAN	24.884 ± 0.991	0.094 ± 0.024	0.881 ± 0.017	24.576 ± 1.073	0.112 ± 0.022	0.852 ± 0.028
	MS-SPADE	25.818 ± 0.857	0.079 ± 0.016	0.904 ± 0.012	25.074 ± 1.085	0.098 ± 0.021	0.867 ± 0.018
	Ours	**30.756** ± 1.950	**0.065** ± 0.034	**0.944** ± 0.031	**32.224** ± 2.518	**0.046** ± 0.029	**0.941** ± 0.025

Table 9: **Modality transfer results of MRI** on BRATS23: The averaged results with standard derivations of metrics between all modalities.

Source	Target	T1			T1ce			T2			Flair		
		PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑	PSNR↑	NMSE↓	SSIM↑
T1	**SwinUNETR**	**32.815**	**0.092**	**0.941**	**31.655**	**0.202**	**0.912**	**24.650**	**0.361**	**0.857**	**27.593**	**0.202**	**0.883**
	Std.	*0.968*	*0.043*	*0.049*	*1.062*	*0.067*	*0.052*	*1.008*	*0.069*	*0.077*	*1.144*	*0.072*	*0.050*
	MS-SPADE	**29.001**	**0.055**	**0.942**	**26.119**	**0.078**	**0.912**	**25.818**	**0.103**	**0.904**	**24.842**	**0.113**	**0.859**
	Std.	*0.643*	*0.025*	*0.022*	*0.816*	*0.022*	*0.015*	*0.857*	*0.030*	*0.014*	*0.728*	*0.034*	*0.019*
	Ours	**43.472**	**0.003**	**0.996**	**34.547**	**0.045**	**0.955**	**30.756**	**0.065**	**0.944**	**31.693**	**0.049**	**0.937**
	Std.	*2.495*	*0.004*	*0.011*	*1.956*	*0.030*	*0.018*	*1.950*	*0.034*	*0.031*	*2.287*	*0.024*	*0.019*
T1ce	**SwinUNETR**	**32.456**	**0.100**	**0.929**	**33.001**	**0.156**	**0.926**	**25.125**	**0.366**	**0.859**	**27.699**	**0.211**	**0.882**
	Std.	*1.018*	*0.044*	*0.048*	*0.889*	*0.055*	*0.051*	*0.964*	*0.071*	*0.074*	*1.129*	*0.071*	*0.049*
	MS-SPADE	**26.228**	**0.076**	**0.922**	**28.759**	**0.060**	**0.937**	**25.990**	**0.092**	**0.907**	**25.204**	**0.092**	**0.881**
	Std.	*0.794*	*0.027*	*0.033*	*0.885*	*0.019*	*0.015*	*0.859*	*0.032*	*0.908*	*0.811*	*0.050*	*0.037*
	Ours	**34.077**	**0.020**	**0.962**	**46.663**	**0.003**	**0.996**	**30.775**	**0.063**	**0.942**	**32.224**	**0.046**	**0.941**
	Std.	*2.484*	*0.012*	*0.017*	*3.240*	*0.004*	*0.008*	*1.812*	*0.030*	*0.028*	*2.518*	*0.029*	*0.025*
T2	**SwinUNETR**	**30.102**	**0.171**	**0.896**	**30.354**	**0.283**	**0.883**	**26.831**	**0.268**	**0.887**	**27.234**	**0.242**	**0.872**
	Std.	*1.405*	*0.056*	*0.050*	*1.249*	*0.086*	*0.054*	*1.144*	*0.054*	*0.075*	*1.154*	*0.073*	*0.051*
	MS-SPADE	**25.422**	**0.085**	**0.908**	**25.234**	**0.087**	**0.895**	**29.230**	**0.048**	**0.942**	**25.074**	**0.098**	**0.867**
	Std.	*0.852*	*0.026*	*0.020*	*1.152*	*0.034*	*0.025*	*0.720*	*0.018*	*0.915*	*1.085*	*0.021*	*0.018*
	Ours	**32.646**	**0.028**	**0.955**	**33.857**	**0.051**	**0.949**	**43.653**	**0.006**	**0.991**	**32.224**	**0.046**	**0.941**
	Std.	*2.391*	*0.028*	*0.028*	*1.925*	*0.040*	*0.027*	*3.467*	*0.024*	*0.038*	*2.518*	*0.029*	*0.025*
Flair	**SwinUNETR**	**31.371**	**0.135**	**0.916**	**31.285**	**0.240**	**0.905**	**25.579**	**0.338**	**0.867**	**29.092**	**0.148**	**0.923**
	Std.	*1.198*	*0.051*	*0.054*	*1.161*	*0.077*	*0.053*	*0.956*	*0.064*	*0.073*	*0.974*	*0.055*	*0.049*
	MS-SPADE	**25.186**	**0.090**	**0.905**	**25.899**	**0.094**	**0.906**	**26.146**	**0.086**	**0.913**	**28.608**	**0.058**	**0.938**
	Std.	*0.759*	*0.028*	*0.048*	*1.039*	*0.025*	*0.027*	*0.636*	*0.028*	*0.944*	*0.769*	*0.025*	*0.028*
	Ours	**32.752**	**0.026**	**0.951**	**33.471**	**0.055**	**0.944**	**30.571**	**0.068**	**0.940**	**43.624**	**0.004**	**0.995**
	Std.	*2.399*	*0.020*	*0.022*	*1.634*	*0.035*	*0.021*	*1.951*	*0.035*	*0.034*	*2.441*	*0.008*	*0.013*

C.2 MISSING MODALITY SEGMENTATION RESULTS ON BRATS18:

We provide detailed segmentation results on BRATS18 as Table 10.

Table 10: **Missing modality segmentation results of MRI** on BRATS18: Num denotes the number of missing modalities for different settings. The used modalities are highlighted with gray boxes and the missing ones remain as blank. The results of each setting are presented accordingly.

Missing Num		=3				=2						=1				=0
Modality	flair															
	T1															
	T1ce															
	T2															
Tumour Core	SPA	65.86	65.27	78.26	66.4	72.99	83.23	70.66	81.25	70.66	80.63	83.22	73.89	83.36	82.05	83.4
	M3AE	69.4	65.45	79.12	71.84	79.9	70.45	82.79	81.17	71.62	73.35	81.78	82.42	73.31	81.61	82.22
	mmFormer	67.8	77.32	64.56	64.08	81.51	79.43	69.14	70.63	68.6	80.75	81.75	70.92	81.74	81.55	82.23
	RFNET	64.03	74.53	58.63	61.95	79.2	77.45	69.25	67.48	67.98	78.85	80.15	70.75	79.4	80.15	80.29
	M2F	65.79	63.29	77.31	63.64	70.38	79.93	68.01	79.62	67.68	79.37	80.65	69.73	80.01	79.53	80.34
	Ours	75.83	71.2	75.29	75.71	80.66	83.6	79.23	74.83	79.51	79.52	83.92	82.78	86.65	81.22	86.72
Enhancing tumour	SPA	39.85	41.39	70.43	41.72	45.99	73.07	45.25	72.87	45.25	72.59	73.52	47.56	73.01	73.55	73.65
	M3AE	37	38.41	75.8	44.22	78.09	45.2	79.36	78.16	41.71	48.12	79.14	80.06	47.63	79.31	79.91
	mmFormer	40.08	72.19	38.89	37.23	73.11	73.06	40.64	42.27	43.65	75.56	43.34	81.74	73.36	75.31	73.4
	RFNET	38.69	69.22	30.89	33.56	71.4	70.9	38.53	41.91	40.9	69.51	71.61	43.37	71.17	74.2	73.79
	M2F	37.99	37.79	71.74	39.28	43.37	74.66	45.42	73.48	43.5	73.48	73.56	45.93	73.15	74.03	75.26
	Ours	67.45	54.83	70.86	47.63	69.38	52.91	70.1	59.45	67.44	63.91	70.79	57.78	69.42	59.76	70.64
Whole tumour	SPA	85.77	72.69	71.95	80.4	87.82	87.97	88.27	75.57	88.27	81.8	88.3	88.78	89.06	82.87	89.27
	M3AE	87.78	74.69	74.91	84.43	76.09	84.48	89.63	84.4	88.64	88.91	84.04	89.29	88.58	88.45	88.26
	mmFormer	84.09	72.85	73.37	85.6	85.97	76.93	87.09	86.09	87.55	87.94	88.36	88.16	88.74	85.96	89.03
	RFNET	80.52	67.06	68.42	82.96	82.57	71.97	85.82	83.25	86	84.94	86.06	86.53	86.34	83.61	86.82
	M2F	85.72	72.48	71.78	82.53	87.73	87.66	84.35	76.03	87.69	84.27	88.17	88.22	88.47	84.32	88.72
	Ours	89.23	81.73	82.26	87.45	89.74	89.03	88.00	81.92	89.72	86.27	89.12	90.5	91.25	87.1	91.19

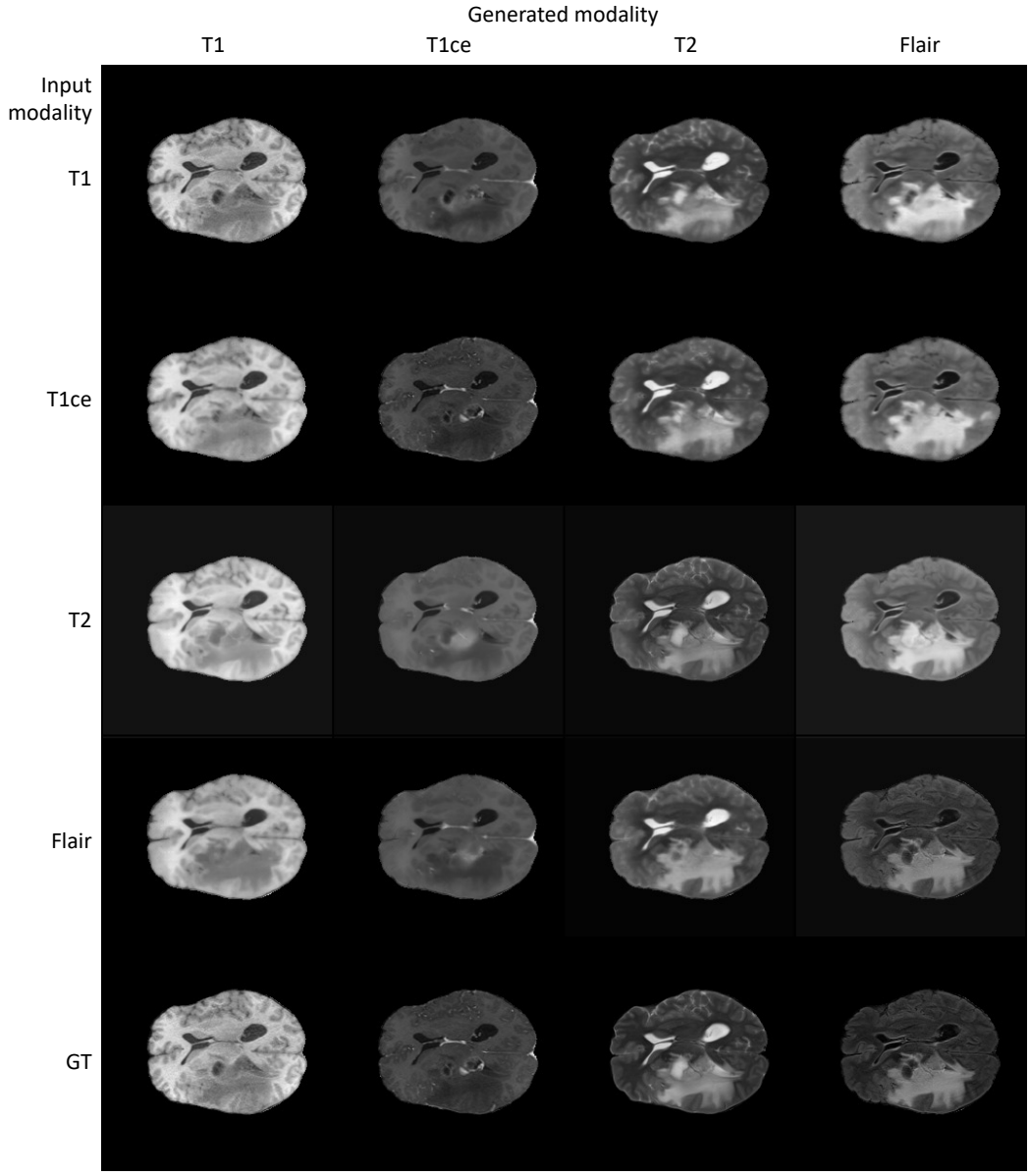

Figure 6: Generated images of our proposed method: slices across ventricles.

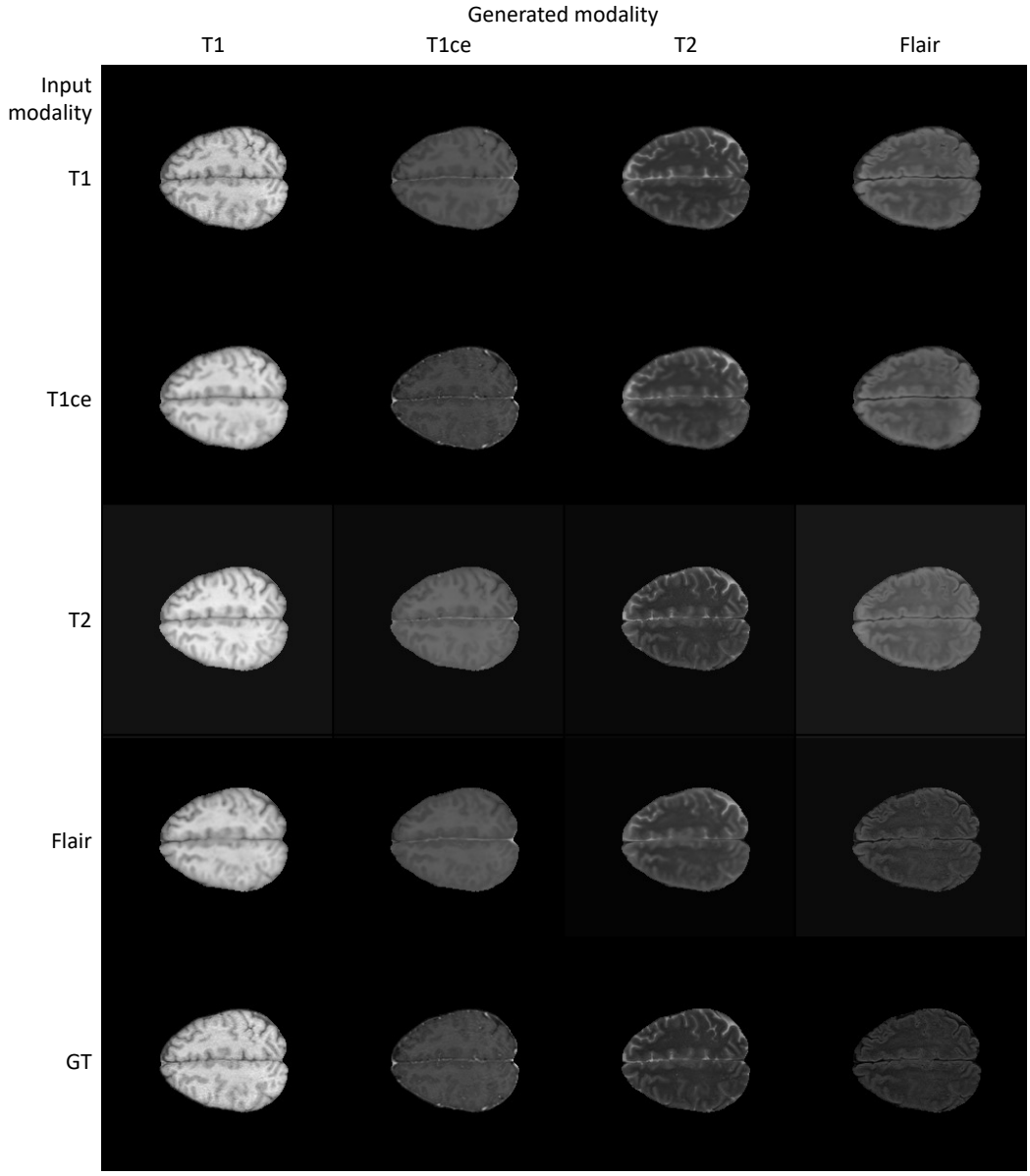

Figure 7: Generated images of our proposed method: slices across cerebral sulcus.

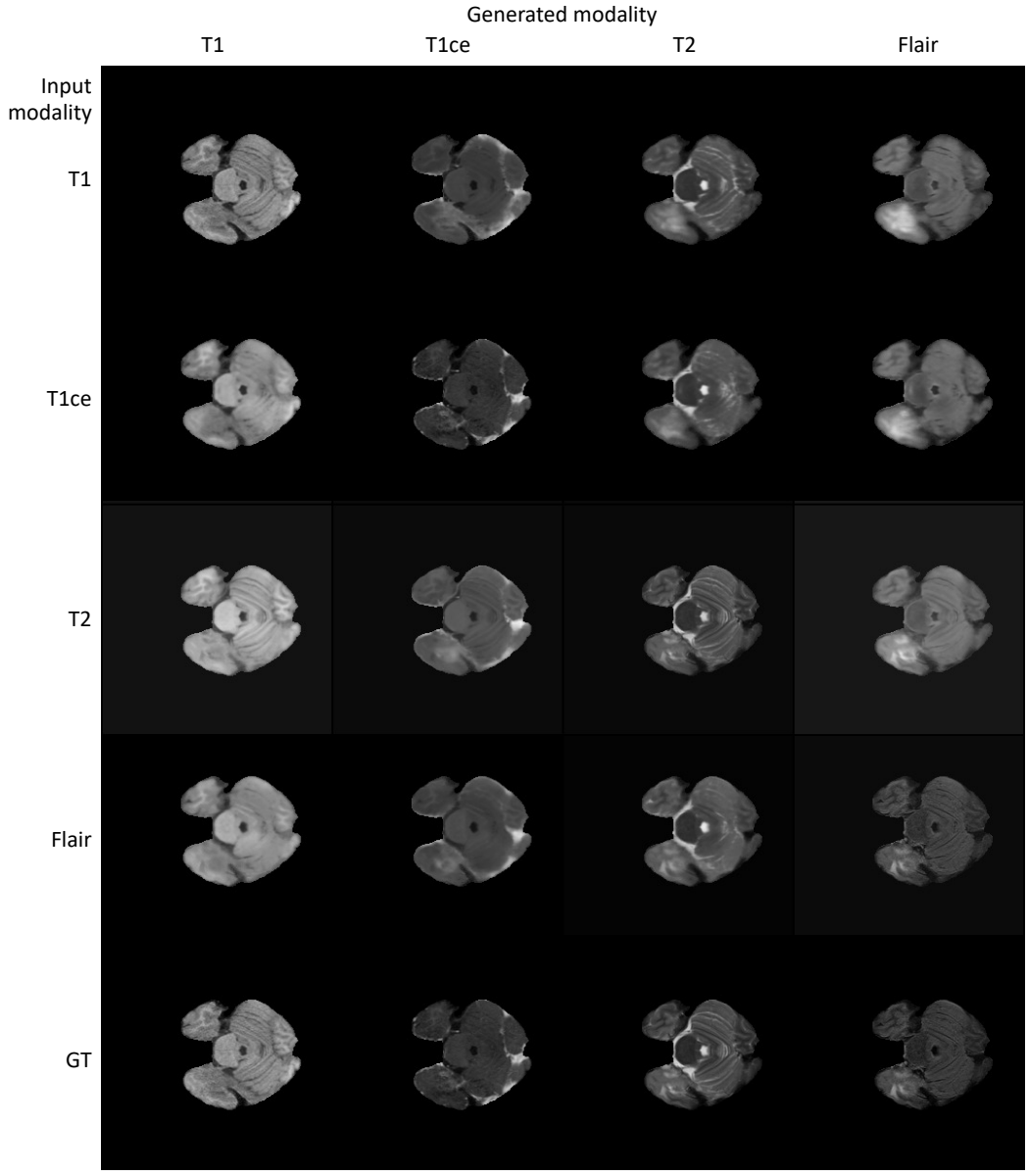

Figure 8: Generated images of our proposed method: slices across cerebellar hemisphere. Our method is able to generate defined cerebellar folia.