Foundation-Guided Representation Alignment for Multimodal Medical Image Registration

Mengjie Guo[1,2], Xinxing Cheng[1], Wenqi Lu[3]*, Qingjie Meng[1], Guanyu Yang[5], Yang Chen[5], Ziyun Ding[1], Alejandro F. Frangi[4], and Jinming Duan[1,4]*

[1] University of Birmingham, Birmingham, UK
[2] Southern University of Science and Technology, Shenzhen, China
[3] Manchester Metropolitan University, Manchester, UK
w.lu@mmu.ac.uk
[4] University of Manchester, Manchester, UK
jinming.duan@manchester.ac.uk
[5] Southeast University, Nanjing, China

Abstract. Multimodal medical image registration is crucial for diagnosis and treatment planning, yet remains challenging due to significant appearance differences between modalities such as MR and CT. The modality gap requires that the encoder in learning-based registration methods has cross-modal capabilities. Meanwhile, recent vision foundation models (VFMs) have demonstrated strong capabilities in extracting rich cross-modal representations that help bridge such modality gaps. Inspired by this, we propose a foundation-guided multimodal representation alignment framework, which transfers the rich cross-modal knowledge from a pretrained VFM to the registration encoder. We introduce a novel Data ReAssembly strategy to transform volumetric medical images into VFM input while preserving spatial information. Additionally, we implement a Hierarchical Representation Fusion module to dynamically integrate multi-level features from the VFM, yielding semantically and structurally rich representations. Extensive experiments on two public MR–CT datasets for brain and abdominal imaging demonstrate the effectiveness of the proposed method compared with existing approaches. The code is available at https://github.com/MeggieGuo/FGRA-Reg.

Keywords: Multimodal image registration · Vision foundation model · Representation learning

1 Introduction

Multimodal medical image registration aims to spatially align images acquired from different imaging modalities, such as magnetic resonance (MR) and computed tomography (CT). Each modality focuses on distinct anatomical or physiological information – for instance, CT provides high-resolution detail of osseous structures, whereas MR offers superior soft tissue contrast. By aligning images

* Corresponding authors.

of different modalities, registration effectively integrates their complementary information, which in turn facilitates clinical interpretation and quantitative analysis. In addition, accurate multimodal alignment is essential for a wide range of downstream tasks, including anatomical atlas mapping, longitudinal patient monitoring, surgical planning, and image fusion. Despite its significance, multimodal image registration remains a challenging problem due to the substantial differences in appearance across modalities [8].

Traditional methods for image registration typically apply iterative optimization algorithms to solve a point-to-point mapping for each pair of images [27]. However, these methods are computationally expensive and tend to be slow in practice. Recent efforts [1,3] are primarily based on deep learning and follow an encoder-decoder paradigm, in which the encoder learns representations essential for identifying anatomical correspondences and the decoder uses these representations to predict the deformation field. However, in multimodal settings, appearance differences across modalities hinder the encoder's ability to learn anatomical correspondences. As a result, equipping the encoder with strong capabilities to extract modality-agnostic and geometry-aware features is crucial, which serves as the key motivation for this work.

Recent vision foundation models (VFMs) [20,21] exhibit strong cross-domain generalizability, thanks to their pretraining on massive datasets and large-scale architectures. These pretrained models capture cross-modal relationships and produce robust and semantically rich features that are essential for a multimodal registration encoder to address inter-modality appearance differences. However, directly applying them to medical image registration faces three primary challenges: (1) Prohibitive computational cost: The high GPU memory demands of VFMs are exacerbated by 3D medical volumes, hindering efficient clinical deployment. (2) Significant domain shift: A fundamental discrepancy exists between the 2D natural images used to pretrain most advanced VFMs and the volumetric medical images. (3) Task objective misalignment: The general-purpose representations of VFMs are not tailored to the specific, localized goal of deformable image registration, complicating direct integration.

To address these challenges, we propose a novel foundation-guided representation alignment framework that enhances multimodal registration by transferring the robust, cross-modal representations of a pretrained VFM to a registration network. We accomplish this through two new components. First, we introduce a Data ReAssembly strategy that converts a volumetric medical image into a 2D format with three channels to meet the input requirements of most VFMs, while effectively preserving crucial 3D spatial information. Additionally, we propose a Hierarchical Representation Fusion module to dynamically integrate features from the VFM at multiple levels, encompassing low-level texture details and high-level semantic information. In experiments, we validate our method on two public 3D MR-CT datasets, and the results demonstrate its superior performance compared to existing state-of-the-art registration methods. Our contributions are summarized as follows:

- We propose a foundation-guided representation alignment framework for multimodal medical image registration, which explicitly reshapes the embedding space of the registration encoder to reduce modality-induced representation gaps.
- We instantiate the framework with two new adaptation modules, Data Re-Assembly to bridge 3D medical volumes and 2D VFM inputs, and Hierarchical Representation Fusion to aggregate multi-level features from the foundation encoder.
- Extensive experiments on multimodal deformable registration benchmarks demonstrate consistent improvements over baselines and validate the generalizability of the proposed framework across different backbones and datasets.

2 Related Work

2.1 Multimodal Medical Image Registration

Multimodal medical image registration is challenging due to substantial modality-dependent intensity variations, which make direct intensity-based registration unreliable. To tackle these challenges, various approaches have been proposed, which we categorize into three groups.

Similarity metric-based methods design loss metrics to measure structural discrepancies across modalities, and minimize the loss to align images. This paradigm enables the adaptation of monomodal registration methods [7, 16, 17, 23, 28, 36] to multimodal tasks simply by replacing the similarity metric, which avoids the need for architectural changes [1, 3, 6, 38]. Classical metrics in this category include Mutual Information (MI) and its variants, which measure statistical dependencies but face challenges in optimization due to insufficient gradient signals and local minima [1, 2]. In addition, several recent works propose learning-based descriptors such as the Modality-Independent Neighbourhood Descriptor (MIND), which can capture local structural patterns but are limited in enforcing global consistency [13].

Image translation-based methods first transform images from one modality into another one, and then perform monomodal registration. For instance, CycleMorph [18] employs a cycle-consistent GAN to enable bidirectional modality translation and facilitate registration, and INNReg [11] proposes a geometry-preserving framework using an Invertible Neural Network to maintain anatomical consistency during translation. However, these approaches suffer from two limitations: first, anatomical artifacts introduced during translation can propagate to the registration stage, resulting in implausible deformation fields; second, most existing methods operate on 2D slices and ignore crucial 3D spatial information in volumetric medical images.

Representation learning-based methods align images by learning a shared, modality-invariant feature space. For example, [29] learns disentangled representations to separate anatomical content from modality-specific appearance information, and SynthMorph [15] learns invariant features from diverse synthetic

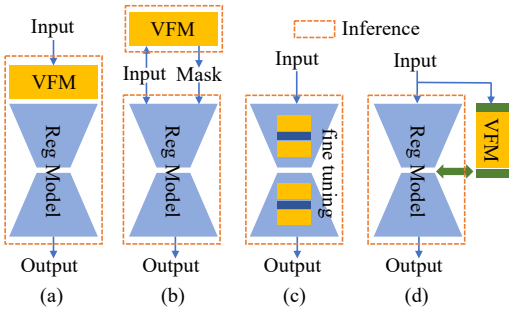

Fig. 1: Mainstream technical frameworks in existing VFM-based medical image processing methods and ours: (a) VFM serves as a feature extractor for data preprocessing; (b) VFM generates auxiliary masks as input; (c) VFM is integrated via adapters for fine-tuning; (d) We transfer cross-modal priors from a VFM. Note that frameworks (a) – (c) rely on VFM during inference, while ours (d) eliminates this dependency.

images for contrast-agnostic registration. However, accurately distinguishing between domain-invariant and domain-specific features is challenging in practice.

2.2 Vision Foundation Models in Medical Imaging

Vision Foundation Models (VFMs) [21] have recently emerged as powerful general-purpose feature extractors for a wide range of downstream tasks. Their adoption in medical image analysis has attracted increasing attention, primarily due to two key advantages: (1) strong transferability and cross-domain generalization enabled by large-scale self-supervised pretraining on massive image datasets, and (2) rich hierarchical representations that capture both fine-grained spatial structures and high-level semantic features.

Recent studies have demonstrated the potential of VFMs in medical image analysis [20], where the mainstream frameworks are summarized in Fig. 1(a)–(c). For example, DINOReg [33,34] applied DINOv2 to extract features as input to the optimization-based multimodal registration method ConvexAdam [31] (Fig. 1(a)), demonstrating the effectiveness of VFM in extracting modality-invariant representations and facilitating multimodal registration. [37] utilized SAM-generated segmentation masks as auxiliary inputs to improve anatomical consistency during registration (Fig. 1(b)). Meanwhile, LLM-Morph [24] embeds pretrained LLM layers into the registration network, so the large model is involved in both training and inference, leading to high computational cost. Moreover, pretrained LLMs are originally optimized for language modeling rather than dense visual correspondence, making additional adaptation necessary for multimodal image registration. (Fig. 1(c)).

Nevertheless, these methods suffer from two key drawbacks. First, they do not adequately address the domain gap when applying VFMs pretrained on 2D natural images to 3D medical volumes. Second, during inference, most approaches require the inclusion of the entire VFM, which hinders computational

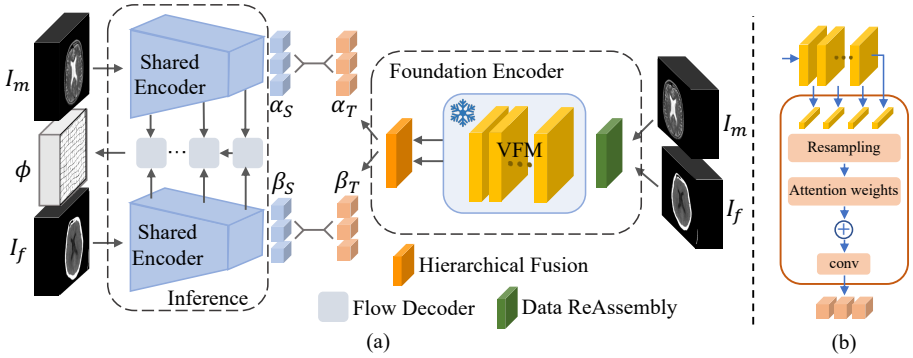

Fig. 2: (a) Overall architecture of the proposed framework for medical image registration. We leverage a pretrained VFM to guide the training of a dual-stream registration network via representation alignment. To bridge the domain gap between volumetric medical images and the 2D natural images generally used in the pretrained VFM, we introduce a Data ReAssembly module. Additionally, a Hierarchical Representation Fusion module (b) dynamically integrates multi-scale features from VFM to enrich the representations transferred to the registration encoder. During inference, only the lightweight registration network is retained.

efficiency and limits clinical applicability. In our proposed framework (Fig. 1(d), see Section 3 for details), we circumvent these limitations and offer an effective strategy for enhancing multimodal image registration.

3 Methodology

In this section, we start by outlining the overall architecture of our foundation-guided registration framework. Then, we detail a Data ReAssembly strategy that effectively bridges the volumetric medical images and the 2D input of VFM. Next, we present a Hierarchical Representation Fusion mechanism to integrate multi-level representations from the VFM. Finally, we introduce the training losses used in the proposed framework.

3.1 Foundation-Guided Registration

Medical image registration aims to find the optimal deformation field, and is typically formulated as

$$\widehat{\phi} = \arg\min_{\phi} \mathcal{L}_{\text{sim}}\left(I_f, I_m \circ \phi\right) + \lambda \mathcal{L}_{\text{smooth}}\left(\phi\right) \tag{1}$$

where I_f and I_m denote the fixed and moving images, respectively, and ϕ represents the deformation field that maps I_m to I_f. The operator $m \circ \phi$ denotes the warped moving image obtained by applying ϕ to I_m. The similarity term

$\mathcal{L}_{\text{sim}}$ measures the alignment quality between the fixed image and the warped image, while $\mathcal{L}_{\text{smooth}}$ acts as a regularization term that encourages smooth and physically plausible deformations.

In this work, we propose a foundation-guided representation alignment framework for multimodal image registration (see Fig. 2). The framework integrates a registration backbone with a pretrained foundation encoder to facilitate cross-modal representation alignment. Specifically, we employ a dual-stream registration backbone consisting of shared encoders and a flow decoder. The shared registration encoder (RE) extracts multi-scale features from the moving and fixed images, denoted as $\alpha_S = \text{RE}(I_m)$ and $\beta_S = \text{RE}(I_f)$, respectively. Based on the encoded features, the flow decoder predicts the deformation field ϕ that warps the moving image toward the fixed image. The specific architectures of the registration encoder and decoder depend on the chosen registration backbone.

To guide representation learning, we further introduce a foundation encoder (FE) that produces cross-modal representations $\alpha_T = \text{FE}(I_m)$ and $\beta_T = \text{FE}(I_f)$. In this work, the FE is instantiated with a pretrained DINOv3 (ViT-B/16) [32], together with two proposed components—Data ReAssembly and Hierarchical Representation Fusion—described in Sections 3.2 and 3.3, respectively. We adopt DINOv3 due to its strong representational capacity learned from large-scale self-supervised pretraining, enabling rich semantic and hierarchical visual representations that generalize well across domains.

We align the representations by minimizing the discrepancy between the registration encoder features (α_S, β_S) and the corresponding features (α_T, β_T) extracted from the foundation encoder. This alignment objective encourages the registration encoder to inherit high-level semantic priors while reducing modality-induced representation gaps, thereby promoting robust and modality-invariant feature learning.

The proposed foundation-guided representation alignment approach enjoys three merits. First, given the limited scale of medical imaging datasets, training a powerful medical foundation model for cross-modal representation from scratch is challenging. Our method overcomes this by transferring the cross-modal representations from a pre-trained VFM to the registration encoder. Second, this framework can augment existing dual-stream registration networks, boosting their performance on multimodal tasks without requiring architectural changes. Third, since only the registration backbone is retained during inference, the proposed framework avoids the computational overhead of VFMs and remains efficient for practical clinical deployment.

3.2 Data ReAssembly

When applying VFMs that are typically pretrained on 2D natural images with three channels $I \in \mathbb{R}^{3\times H\times W}$ to 3D medical image processing tasks, it is necessary to reformat the volumetric medical images $V \in \mathbb{R}^{H\times W\times D}$. Fig. 3(a) describes a common way [10, 33], which extracts 2D slices from the 3D volume and repeats each slice three times to create a three-channel input. However, this repetition leads to significant data redundancy and results in inputs that substantially differ

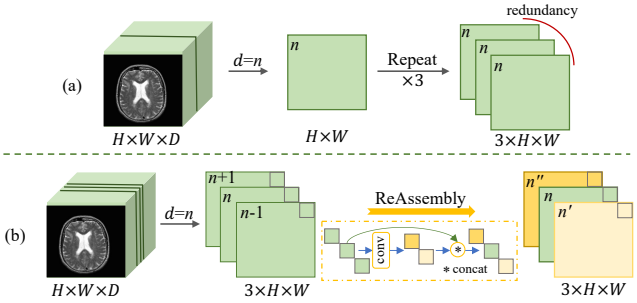

Fig. 3: Illustration of the proposed Data ReAssembly strategy. Existing approaches typically form 2D inputs by extracting a single slice from a volumetric medical image and repeating it across three channels (a). In contrast, our method extracts multiple adjacent slices and processes them via a learnable reassembly module to construct an informative, non-redundant three-channel representation (b).

from the natural images used in the pretraining VFM. Besides, this approach discards the potentially useful spatial information in the 3D volume.

To avoid redundancy and preserve spatial information, we introduce a Data ReAssembly (DR) strategy. As illustrated in Fig. 3(b), given a 2D slice S_n extracted from a 3D volume $\nu \in \mathbb{R}^{H\times W\times D}$ at depth $d = n, n \in [0, D)$, we first gather its adjacent slices S_{n-1} and S_{n+1} to form a 2.5D tensor [39]:

$$V_{2.5D} = [S_{n-1}, S_n, S_{n+1}] \tag{2}$$

Although this 2.5D tensor already incorporates local spatial context, using it directly as input to a pretrained VFM is suboptimal. The three channels in $V_{2.5D}$ encode neighboring spatial slices rather than the color semantics expected by VFM pretraining, leading to a distribution mismatch between 2.5D inputs and natural image statistics. To overcome this, a lightweight context fusion module, implemented as a 1×1 convolution, is introduced to map $V_{2.5D}$ into two informative feature channels:

$$S_{n'}, S_{n''} = \mathrm{Conv}(V_{2.5D}) \tag{3}$$

These two synthesized channels, together with the original slice S_n, form a new non-redundant three-channel input:

$$\tilde{I} = [S_{n'},\ S_n,\ S_{n''}] \tag{4}$$

This reassembled input $\tilde{I}$ effectively encodes local structural context from the 3D volume while remaining compatible with 2D pretrained VFM, enabling more spatially consistent feature learning for registration tasks.

The DR strategy offers three key benefits: (1) it eliminates the redundancy caused by simple slice repetition, (2) it preserves limited but valuable spatial context in the 3D volume, and (3) the learnable reassembly allows the input

data to be adaptively transformed, thereby better aligning the medical imaging domain with VFMs. It is worth noting that DR is applied only in the foundation encoder branch, while the shared registration encoders still take the full 3D volumes as input; therefore, the predicted deformation field is generated by the 3D registration backbone and preserves volumetric consistency.

3.3 Hierarchical Representation Fusion

The VFM can usually generate a hierarchy of representations, where earlier layers tend to preserve high-resolution, fine-grained spatial information, while deeper layers produce more refined, abstract semantic embeddings. Both localized details of early features and the holistic understanding of deep features are crucial for establishing dense correspondences in registration. However, to the best of our knowledge, current methods [5,22,33] that utilize VFMs tend to overlook this hierarchical nature, and they generally rely solely on the final layer's output for downstream tasks.

To leverage these multi-level features holistically, we propose a Hierarchical Representation Fusion (HRF) module, illustrated in Fig. 2(b). The HRF module first extracts the multi-level features $\mathbf{F} = \{F_1, F_2, ..., F_L\}$ from the VFM encoder. Each feature F_l, $l \leq L$ is generated from one of the encoder's layers, and is then resampled to align with the 3D spatial dimensions:

$$\tilde{F}_l = \text{Resample}(F_l; H, W, D), F_l \in \mathbf{F} \tag{5}$$

where $H \times W \times D$ is the target spatial dimension and is aligned with that of the registration encoder's representation. Second, the HRF module introduces attention weights to fuse the multi-level features dynamically:

$$\tilde{F}_{\text{fused}} = \sum_{l=1}^{L} \frac{\exp(w_l)}{\sum_{j=1}^{L} \exp(w_j)} \tilde{F}_l \tag{6}$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_L) \in \mathbb{R}^L$ is the learnable weight for highlighting useful features. Third, a convolutional layer transforms the fused features to ensure that the outputs, α_T and β_T, have the same channels as α_S and β_S. This operation is crucial for this framework to flexibly adapt to various registration backbones.

3.4 Total Loss

The total loss consists of three components: a representation alignment loss, a similarity loss, and a regularization loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{sim}} + \mathcal{L}_{\text{smooth}} \tag{7}$$

- *Representation alignment loss* $\mathcal{L}_{rep}$: We employ feature-level supervision to align the dense feature maps produced by the registration encoder with those

from the foundation encoder. The loss is defined as the sum of cosine similarity losses between corresponding feature pairs:

$$\mathcal{L}_{\text{rep}} = d(\alpha_S, \alpha_T) + d(\beta_S, \beta_T) \tag{8}$$

$$d(\mathbf{x_1}, \mathbf{x_2}) = 1 - \frac{\sum_i x_1^i x_2^i}{\sqrt{\sum_i (x_1^i)^2}\sqrt{\sum_i (x_2^i)^2}} \tag{9}$$

where α_S and α_T denote the feature maps of the moving image from the registration and foundation encoders RE and FE, respectively. Similarly, β_S and β_T are feature maps of the fixed image.

- *Similarity loss* $\mathcal{L}_{sim}$: It is designed based on the mutual information (MI) loss [30] and the Dice loss (*non-mandatory*) [1]:

$$\mathcal{L}_{\text{sim}} = \mathcal{L}_{\text{MI}}(I_f, I_m \circ \phi) + \mathcal{L}_{\text{Dice}}(S_f, S_m \circ \phi)_{\text{optional}} \tag{10}$$

where ϕ denotes the predicted deformation field, I_f/I_m and S_f/S_m represent the fixed/moving images and their corresponding segmentation labels, $\mathcal{L}_{\text{MI}}$ and $\mathcal{L}_{\text{Dice}}$ represent the MI loss and the Dice loss, respectively. The MI loss is widely used in cross-modality registration due to its effectiveness in capturing statistical dependencies between image intensities. The Dice loss is optional and can be incorporated using auxiliary segmentation labels to improve anatomical alignment during training, whereas segmentation labels are not required during inference.

- *Regularization loss* $\mathcal{L}_{smooth}$*:* This loss enforces spatial smoothness and structural plausibility, and takes the following form:

$$\mathcal{L}_{\text{smooth}}(\phi) = \sum_{\mathbf{p} \in \Omega} \|\nabla \mathbf{u}(\mathbf{p})\|^2 \tag{11}$$

where $\mathbf{p}$ is a spatial coordinate (e.g., a voxel in 3D or a pixel in 2D) within the image domain Ω and $\phi(\mathbf{p}) = \mathbf{p} + \mathbf{u}(\mathbf{p})$ denotes the deformation field.

4 Experiments

4.1 Datasets

We evaluated our method on two multimodal medical image datasets, including abdomen MR-CT data from the Learn2Reg dataset [14] and brain MR-CT data from the SR-Reg dataset [12, 35].

The Learn2Reg Dataset The dataset consists of 40 MR volumes and 49 CT volumes that are unpaired, along with 8 paired MR-CT volumes. Each volume has the dimension of $192 \times 160 \times 192$, and segmentation labels for the liver, spleen, and left/right kidneys are provided for all volumes. The unpaired MR and CT volumes are randomly matched in training and validation, while the 8

Table 1: Quantitative results on the Learn2Reg dataset. Mean±standard deviation are reported for DSC (%), HD95, and $\%(|J_\phi| \leq 0)$.

Method	DSC/%	HD95/mm	$\%(\lvert J_\phi\rvert \leq 0)$	Params/M
Initial	57.24±19.42	11.88±7.44	-	-
NiftyReg	72.61±22.89	8.54±8.34	**0.00±0.00**	-
LapIRN	70.91±19.30	10.48±6.05	0.10±0.01	1.20
VoxelMorph	67.81±18.50	10.23±6.24	0.02±0.01	**0.30**
TransMorph	70.97±20.68	10.64±7.96	0.07±0.00	46.76
TransMatch	70.45±20.74	11.12±8.96	0.02±0.01	112.38
UTSRMorph	75.22±18.70	9.72±7.77	0.03±0.01	98.82
PiViT	67.15±24.28	11.55±8.26	0.01±0.00	0.66
Ours$_\mathbf{P}$	71.26±18.82	10.47±7.61	0.01±0.00	0.66
ModeT	73.84±23.64	7.80±6.22	0.02±0.02	1.03
Ours$_\mathbf{M}$	**76.16±20.93**	**7.12±5.74**	0.03±0.02	1.03

paired volumes are used for testing. We affinely align the MR image with the CT image during preprocessing [3]. In all experiments, the MR volumes serve as the moving images and the CT volumes as the fixed images. The data splitting and image preprocessing remain consistent across all experiments, and all image intensities are normalized to $[0, 1]$.

The SR-Reg Dataset The dataset is introduced as a part of the SynthRAD 2023 dataset [35], and reprocessed by [12] for deformable registration tasks. It consists of 180 skull-stripped, intensity-normalized, and anatomically aligned MR-CT image pairs, each with corresponding segmentation labels. We split the dataset into 150/10/20 subjects for training, validation, and testing, respectively. All volumes are center-cropped to the size of $160 \times 192 \times 192$ with an isotropic resolution of $1 \times 1 \times 1\ mm^3$. Moreover, since the MR-CT image pairs in SR-Reg are aligned, we apply synthetic deformations to the MR images to generate misaligned image pairs for deformable registration [9].

4.2 Evaluation Metrics

Anatomical Accuracy To quantify the alignment of anatomical structures, we use the Dice Similarity Coefficient (DSC) and the 95th percentile Hausdorff Distance (HD95). In particular, DSC measures the volumetric overlap between corresponding labels in the warped and fixed images, while HD95 measures the 95th percentile surface distance between them, providing a robust assessment of boundary alignment.

Deformation Field Quality We quantify the regularity of the deformation fields by the percentage of voxels with non-positive Jacobian determinants ($\%(|J_\phi| \leq 0)$), where smaller values indicate better regularity.

Table 2: Quantitative evaluation results on the SR-Reg dataset.

Methods	DSC/%	HD95	%($\lvert J_\phi\rvert \leq 0$)	MSE/%	SSIM/%	PSNR	Params/M
Initial	55.14±4.81	4.33±0.55	-	2.65±0.25	66.58±2.07	33.91±0.28	-
VoxelMorph	69.79±4.89	2.92±0.56	**0.01±0.00**	0.97±0.15	76.65±2.06	34.07±0.33	**0.30**
TransMorph	71.16±5.24	2.96±0.67	0.49±0.13	0.99±0.19	77.15±2.04	34.03±0.34	46.76
UTSRMorph	75.20±4.57	2.32±0.58	0.05±0.05	0.85±0.13	78.53±2.31	34.16±0.38	98.82
ModeT	72.30±3.63	2.58±0.41	0.01±0.01	0.88±0.21	78.05±2.00	34.15±0.33	1.03
Ours$_{\mathrm{M}}$	**79.04±2.37**	**1.97±0.30**	0.06±0.02	**0.83±0.20**	**80.00±2.27**	**34.19±0.41**	1.03

Image Similarity on Synthetic Data A key advantage of using datasets with synthetic deformations is the availability of the ground truth image, which allows for a direct evaluation of registration fidelity beyond anatomical labels. On the SR-Reg dataset, we compute a suite of image similarity metrics — Mean Squared Error (MSE), Structural Similarity Index Measure (SSIM), and Peak Signal-to-Noise Ratio (PSNR) — between the warped moving images and the ground-truth images.

Computational Cost To provide a holistic comparison, we report the number of model parameters in each method. These statistics quantitatively illustrate the trade-off between registration accuracy and computational cost, which is critical for practical clinical deployment.

4.3 Implementation Details

All experiments are implemented in PyTorch 2.8 and Python 3.11 on a single NVIDIA H100 GPU (80GB). We use the Adam optimizer [19] with a batch size of 1 and an initial learning rate of 2×10^{-4}. Models are trained for 100 epochs on the SR-Reg dataset and 400 epochs on the Learn2Reg dataset. To enhance registration accuracy, we incorporate segmentation masks as auxiliary supervision for all learning-based methods during training. The HRF module extracts features from layers 3, 6, 9, and 12 of DINOv3's 12-layer ViT encoder, selected based on preliminary experiments to capture effective multi-level representations.

4.4 Experimental Results

To validate the effectiveness of the proposed method, we compare it with six deformable registration methods, including a traditional method, NiftyReg [27], and five learning-based approaches: LapIRN [28], ViT-V-Net [4], VoxelMorph [1], TransMorph [3], TransMatch [6], and UTSRMorph [38]. We utilize ModeT [36] as our registration backbone.

Results on the Learn2Reg Dataset Registration of the abdomen MR-CT images from the Learn2Reg dataset is a particularly challenging task due to substantial anatomical deformations and the very limited number of paired samples. We first evaluate the proposed framework built on ModeT (Ours$_{\mathrm{M}}$), and further

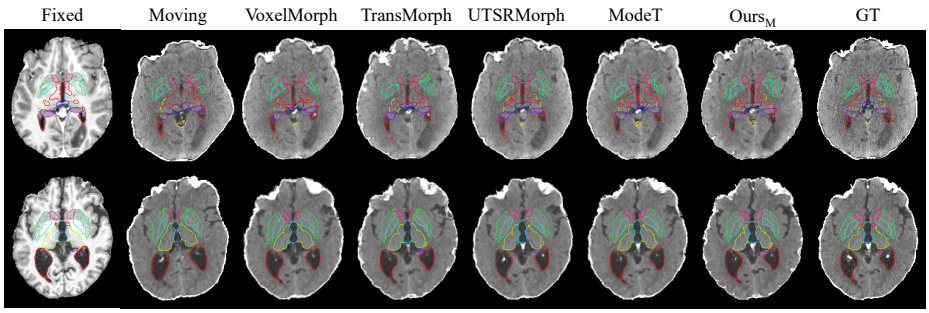

Fig. 4: Visualized registration results from different methods on the SR-Reg dataset. The two rows are registered images with segmentation overlays from two samples.

assess its generalizability by integrating it into another dual-stream registration backbone, PiViT [25], resulting in Ours$_P$. The quantitative results are summarized in Table 1. The proposed method Ours$_M$ achieves a DSC of 76.16% and an HD95 of 7.12 mm, outperforming the original ModeT as well as all other competing methods. In addition, Ours$_P$ consistently improves both DSC and HD95 compared with the original PiViT and most baseline methods while requiring considerably fewer parameters.

These results demonstrate the effectiveness of the proposed framework in achieving accurate anatomical alignment under severe modality discrepancies and large anatomical deformations, while also highlighting its flexibility to be integrated with different registration backbones.

Results on the SR-Reg Dataset We evaluate the proposed framework Ours$_M$ built upon ModeT. The comparison covers six quantitative metrics reflecting anatomical accuracy, structural similarity, and deformation regularity.

As shown in Table 2, the proposed framework improves the performance of the registration backbone. Specifically, compared with ModeT, Ours$_M$ achieves a substantial 6.74% improvement in DSC and a notable reduction in HD95 (from 2.58 to 1.97), indicating significantly enhanced anatomical alignment accuracy. Moreover, these gains are achieved while preserving competitive image similarity metrics (MSE, SSIM, and PSNR) and retaining a lightweight registration backbone. These results demonstrate that the proposed framework enhances multimodal image registration on brain MR-CT data. Qualitative comparisons in Fig. 4 further corroborate these findings, showing improved structural alignment under our framework.

Representation Visualization To further investigate the effectiveness of the proposed framework in promoting cross-modal representation alignment, we visualize the learned features using t-SNE [26]. Specifically, we extract registration encoder features from both ModeT (denoted as Base) and our framework (de-

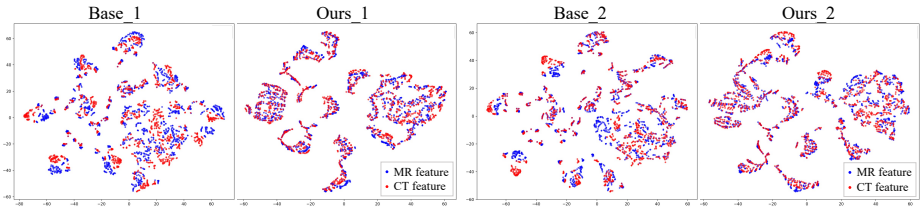

Fig. 5: Visualizing cross-modal representations by t-SNE.

noted as Ours) for two samples from the SR-Reg test set. For each sample, we project the MR and CT features into a 2D embedding space using t-SNE.

As shown in Fig. 5, compared with Base_1 and Base_2, our framework (Ours_1 and Ours_2) demonstrates significantly improved cross-modal alignment at the feature level. Specifically, MR and CT features from similar anatomical regions are more closely aligned and form compact clusters with clearer boundaries, which suggests enhanced structural discriminability and reduced modality-induced distribution gaps. These observations provide qualitative evidence that our foundation-guided representation alignment mechanism effectively encourages modality-invariant feature learning, thereby facilitating more reliable multimodal deformable registration.

Cross-dataset Generalization To evaluate the generalization capability of our framework, we conduct cross-dataset transfer experiments from the Learn2Reg abdomen dataset to the SR-Reg brain dataset. Specifically, we train ModeT [36] and PiViT [25], along with their foundation-guided variants (Ours_{M} and Ours_{P}), on the Learn2Reg abdomen dataset.

For transfer to the SR-Reg brain dataset, we randomly select 20 samples from the SR-Reg training set for fine-tuning. To evaluate the generalization capability of the learned representations, the foundation encoder is removed, and the pretrained registration encoder is kept frozen, while only the decoder parameters are updated. The models are fine-tuned using the Adam optimizer with a learning rate of 1×10^{-4} for 150 epochs, and subsequently evaluated on the SR-Reg brain testing set.

The quantitative results are summarized in Table 3. Both Ours_{M} and Ours_{P} consistently outperform their corresponding baselines, ModeT and PiViT. In particular, the proposed framework achieves higher DSC and lower HD95 while maintaining comparable topology preservation. Since the registration encoder remains frozen during transfer, the observed improvements suggest that our framework enables it to learn more transferable and robust features across datasets with varying anatomical structures and imaging distributions.

Table 3: Cross-dataset generalization.

Methods	DSC/%	HD95	$\%(\lvert J_\phi\rvert \leq 0)$
ModeT	62.89±6.10	3.89±0.79	0.00±0.00
Ours$_{\mathbf{M}}$	64.83±5.02	3.64±0.68	0.00±0.00
PiViT	66.83±4.47	3.33±0.72	0.05±0.05
Ours$_{\mathbf{P}}$	68.36±5.15	3.11±0.80	0.03±0.03

Table 4: Ablation study.

Methods	DSC/%	HD95	$\%(\lvert J_\phi\rvert \leq 0)$
Base (ModeT)	72.30±3.63	2.58±0.41	0.01±0.01
+RA	76.83±2.73	2.23±0.37	0.04±0.02
+RA+DR	77.26±2.89	2.16±0.37	0.07±0.03
+RA+HRF	78.08±2.47	2.09±0.31	0.08±0.03
Ours	79.04±2.37	1.97±0.30	0.06±0.02

4.5 Ablation Study

We conduct ablation experiments on the SR-Reg dataset to assess the contribution of each component in our proposed framework. The quantitative results are summarized in Table 4.

Representation Alignment (RA). This experiment evaluates the effect of our RA strategy by comparing ModeT and ModeT with RA. In ModeT, following common practice in prior works, we constructed the input to VFM by extracting 2D slices from 3D volumes and repeating each slice three times to form a three-channel input. Features from the final layer of the VFM encoder are processed via a convolutional layer to align spatially with the representations of the ModeT encoder. These features are then aligned using the loss $\mathcal{L}_{\text{rep}}$ (Eq. (8)). By Table 4, RA alone yields a notable improvement in registration performance.

Data ReAssembly (DR). In this setup, we replace the repeated-slice input strategy used in the RA experiment with our proposed DR strategy. This modification results in additional performance gains, underscoring the importance of utilizing structured and non-redundant input representations.

Hierarchical Representation Fusion (HRF). Instead of relying solely on the final-layer output, we employ the HRF module to integrate multi-level features from the VFM encoder. This approach leads to a considerable improvement over the above RA setting, demonstrating the value of the HRF module.

Full Model. Finally, we include all three components — RA, DR, and HRF — to form the complete proposed model. From Table 4, the full model outperforms all ablated variants, which validates the complementary role of each component. Further ablations and experiments are provided in the supplementary material.

5 Discussion

Multimodal deformable registration fundamentally requires bridging modality-induced domain gaps. While existing approaches mainly focus on designing modality-robust similarity metrics or integrating handcrafted descriptors, our framework reshapes the embedding space of the registration encoder under the guidance of a pretrained foundation encoder. This alignment encourages cross-modal consistency while preserving structural discriminability, thereby stabilizing deformation field optimization.

Despite its effectiveness, the performance of our framework depends on the cross-domain generalization capability of the adopted foundation models. Although we employ DINOv3 in our experiments due to its large-scale self-supervised pretraining on massive natural image datasets and strong semantic representation capacity, natural image statistics still differ from medical imaging characteristics. In contrast, current medical vision foundation models and 3D large models are typically trained on comparatively limited datasets and remain limited in scale and diversity. As a result, there exists a trade-off between representational richness and domain specificity when selecting a foundation model.

In addition, our framework may inherit limitations from the foundation model. If the foundation model provides unreliable representations for certain anatomical structures, the alignment loss may transfer suboptimal or biased features to the registration encoder. This risk is partially mitigated because the foundation encoder only provides auxiliary representation guidance during training, while the deformation field is predicted by the registration backbone and constrained by the image similarity and deformation regularization losses, rather than being directly estimated from foundation-model features [33, 34].

Future work will investigate adaptive foundation model selection strategies and explore the integration of emerging large-scale medical or volumetric foundation models. Such directions may further enhance cross-modal representation alignment and improve generalization in complex 3D medical imaging scenarios.

6 Conclusion

In this work, we proposed a foundation-guided cross-modal representation alignment framework for multimodal medical image registration. By leveraging pretrained VFMs, the proposed approach transfers rich cross-modal features to guide the representation learning of the registration encoder, thereby reducing modality-induced representation gaps. To effectively integrate foundation representations, we introduced two key components: Data ReAssembly, which bridges the discrepancy between volumetric medical images and the 2D inputs used in VFM pretraining, and Hierarchical Representation Fusion, which adaptively aggregates multi-level features. Extensive experiments on two publicly available 3D MR–CT datasets demonstrate that the proposed framework achieves superior registration performance compared with state-of-the-art methods. Furthermore, the framework consistently improves different registration backbones and exhibits strong cross-dataset generalization capability, highlighting the effectiveness and flexibility of foundation-guided representation alignment for multimodal medical image registration.

Acknowledgements

The computations were performed using Baskerville, funded by the EPSRC and UKRI (EP/T022221/1 and EP/W032244/1) and operated at the University of Birmingham.

References

1. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. IEEE Transactions on Medical Imaging **38**(8), 1788–1800 (2019)
2. Cao, X., Yang, J., Wang, L., Xue, Z., Wang, Q., Shen, D.: Deep learning based inter-modality image registration supervised by intra-modality similarity. In: International Workshop on Machine Learning in Medical Imaging. pp. 55–63. Springer (2018)
3. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. Medical Image Analysis **82**, 102615 (2022)
4. Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y.: Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. arXiv preprint arXiv:2104.06468 (2021)
5. Chen, K.: Brain imaging foundation model with DINOv2 for image registration. Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. CS231n (2024)
6. Chen, Z., Zheng, Y., Gee, J.C.: Transmatch: A transformer-based multilevel dual-stream feature matching network for unsupervised deformable image registration. IEEE Transactions on Medical Imaging **43**(1), 15–27 (2023)
7. Cheng, X., Miller, T., Lu, W., Meng, Q., Frangi, A.F., Duan, J.: Sacb-net: Spatial-awareness convolutions for medical image registration. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5227–5237 (2025), https://api.semanticscholar.org/CorpusID:277313165
8. Darzi, F., Bocklitz, T.: A review of medical image registration for different modalities. Bioengineering **11**(8), 786 (2024)
9. Deng, X., Liu, E., Li, S., Duan, Y., Xu, M.: Interpretable multi-modal image registration network based on disentangled convolutional sparse coding. IEEE Transactions on Image Processing **32**, 1078–1091 (2023)
10. Dey, N., Billot, B., Wong, H.E., Wang, C.J., Ren, M., Grant, P.E., Dalca, A.V., Golland, P.: Learning general-purpose biomedical volume representations using randomized synthesis (2024), https://arxiv.org/abs/2411.02372
11. Guo, M.: Unsupervised multi-modal medical image registration via invertible translation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 22–38. Springer Nature Switzerland, Cham (2025)
12. Guo, T., Wang, Y., Meng, C.: Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration. CoRR (2024)
13. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical Image Analysis **16**(7), 1423–1435 (2012)
14. Hering, A., Hansen, L., Mok, T.C., Chung, A.C., Siebert, H., Häger, S., Lange, A., Kuckertz, S., Heldmann, S., Shao, W., et al.: Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. IEEE Transactions on Medical Imaging **42**(3), 697–712 (2022)
15. Hoffmann, M., Billot, B., Greve, D.N., Iglesias, J.E., Fischl, B., Dalca, A.V.: Synthmorph: Learning contrast-invariant registration without acquired images. IEEE Transactions on Medical Imaging **41**(3), 543–558 (2022). https://doi.org/10.1109/TMI.2021.3116879

16. Hu, B., Zhou, S., Xiong, Z., Wu, F.: Recursive decomposition network for deformable image registration. IEEE Journal of Biomedical and Health Informatics **26**(10), 5130–5141 (2022). https://doi.org/10.1109/JBHI.2022.3189696
17. Kang, M., Hu, X., Huang, W., Scott, M.R., Reyes, M.: Dual-stream pyramid registration network. Medical Image Analysis **78**, 102379 (2022)
18. Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C.: Cyclemorph: cycle consistent unsupervised deformable image registration. Medical Image Analysis **71**, 102036 (2021)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Liang, P., Pu, B., Huang, H., Li, Y., Wang, H., Ma, W., Chang, Q.: Vision foundation models in medical image analysis: Advances and challenges. arXiv preprint arXiv:2502.14584 (2025)
21. Lin, F.: Vision language models: A survey of 26k papers. arXiv preprint arXiv:2510.09586 (2025)
22. Liu, C., Chen, Y., Shi, H., Lu, J., Jian, B., Pan, J., Cai, L., Wang, J., Zhang, Y., Li, J., et al.: Does DINOv3 set a new medical vision standard? arXiv preprint arXiv:2509.06467 (2025)
23. Liu, Y., Zuo, L., Han, S., Xue, Y., Prince, J.L., Carass, A.: Coordinate translator for learning deformable medical image registration. In: International Workshop on Multiscale Multimodal Medical Imaging. pp. 98–109. Springer (2022)
24. Ma, M., Wang, W., Ning, J., He, J., Sebe, N., Lepri, B.: Large language models for multimodal deformable image registration. arXiv preprint arXiv:2408.10703 (2024)
25. Ma, T., Dai, X., Zhang, S., Wen, Y.: Pivit: Large deformation image registration with pyramid-iterative vision transformer. In: Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., Taylor, R. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2023. pp. 602–612. Springer Nature Switzerland, Cham (2023)
26. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(86), 2579–2605 (2008), http://jmlr.org/papers/v9/vandermaaten08a.html
27. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Computer Methods and Programs in Biomedicine **98**(3), 278–284 (2010)
28. Mok, T.C., Chung, A.C.: Large deformation diffeomorphic image registration with laplacian pyramid networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 211–221. Springer (2020)
29. Qin, C., Shi, B., Liao, R., Mansi, T., Rueckert, D., Kamen, A.: Unsupervised deformable registration for multi-modal images via disentangled representations. In: International Conference on Information Processing in Medical Imaging. pp. 249–261. Springer (2019)
30. Sengupta, D., Gupta, P., Biswas, A.: A survey on mutual information based medical image registration algorithms. Neurocomputing **486**, 174–188 (2022)
31. Siebert, H., Großbröhmer, C., Hansen, L., Heinrich, M.P.: Convexadam: Self-configuring dual-optimisation-based 3d multitask medical image registration. IEEE Transactions on Medical Imaging (2024)
32. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)

33. Song, X., Xu, X., Yan, P.: Dino-reg: General purpose image encoder for training-free multi-modal deformable medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 608–617. Springer (2024)
34. Song, X., Xu, X., Zhang, J., Machado Reyes, D., Yan, P.: Dino-reg: Efficient multimodal image registration with distilled features. IEEE Transactions on Medical Imaging **44**(9), 3809–3819 (2025). https://doi.org/10.1109/TMI.2025.3567247
35. Thummerer, A., Van der Bijl, E., Galapon Jr, A., Verhoeff, J.J., Langendijk, J.A., Both, S., van den Berg, C.N.A., Maspero, M.: Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. Medical Physics **50**(7), 4664–4674 (2023)
36. Wang, H., Ni, D., Wang, Y.: Modet: Learning deformable image registration via motion decomposition transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 740–749. Springer (2023)
37. Xu, H., Xue, T., Fan, J., Liu, D., Chen, Y., Zhang, F., Westin, C.F., Kikinis, R., O'Donnell, L.J., Cai, W.: Medical image registration meets vision foundation model: Prototype learning and contour awareness. arXiv preprint arXiv:2502.11440 (2025)
38. Zhang, R., Mo, H., Wang, J., Jie, B., He, Y., Jin, N., Zhu, L.: Utsrmorph: a unified transformer and superresolution network for unsupervised medical image registration. IEEE Transactions on Medical Imaging (2024)
39. Zhang, Y., Liao, Q., Ding, L., Zhang, J.: Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5 d solutions. Computerized Medical Imaging and Graphics **99**, 102088 (2022)