



Disentangling Progress in Medical Image Registration: Beyond Trend-Driven Architectures towards Domain-Specific Strategies

Bailiang Jian^{♥a,b}, Jiazhen Pan^{♣a}, Rohit Jena^d, Morteza Ghahremani^{ia,b}, Hongwei Bran Li^{ia,e}, Daniel Rueckert^{a,b,c}, Christian Wachinger^{♣a,b}, Benedikt Wiestler^{♣a,b}

^aTechnical University of Munich, Munich, Germany

^bMunich Center for Machine Learning (MCML), Munich, Germany

^cImperial College London, London, England

^dUniversity of Pennsylvania, Pennsylvania, USA

^eNational University of Singapore, Singapore

ARTICLE INFO

Article history:

Keywords: Medical Image Registration, Architectural Design, Open Benchmark

ABSTRACT

Medical image registration drives quantitative analysis across organs, modalities, and patient populations. Recent deep learning methods often combine low-level “trend-driven” computational blocks from computer vision, such as large-kernel CNNs, Transformers, and state-space models, with high-level registration-specific designs like motion pyramids, correlation layers, and iterative refinement. Yet, their relative contributions remain unclear and entangled. This raises a central question: should future advances in registration focus on importing generic architectural trends or on refining domain-specific design principles? Through a modular framework spanning brain, lung, cardiac, and abdominal registration, we systematically disentangle the influence of these two paradigms. Our evaluation reveals that low-level “trend-driven” computational blocks offer only marginal or inconsistent gains, while high-level registration-specific designs consistently deliver more accurate, smoother, and more robust deformations. These domain priors significantly elevate the performance of a standard U-Net baseline, far more than variants incorporating “trend-driven” blocks, achieving an average relative improvement of $\sim 3\%$. All models and experiments are released within a transparent, modular benchmark that enables plug-and-play comparison for new architectures and registration tasks (<https://github.com/BailiangJ/rethink-reg>). This dynamic and extensible platform establishes a common ground for reproducible and fair evaluation, inviting the community to isolate genuine methodological contributions from domain priors. Our findings advocate a shift in research emphasis: from following architectural trends to embracing domain-specific design principles as the true drivers of progress in learning-based medical image registration.

© 2025 Elsevier B. V. All rights reserved.

Preprint under review at Medical Image Analysis

♣ Equal Advising.

♣ Corresponding Author.

♥ Author Contact: {bailiang.jian, jiazhen.pan}@tum.de

1. Introduction

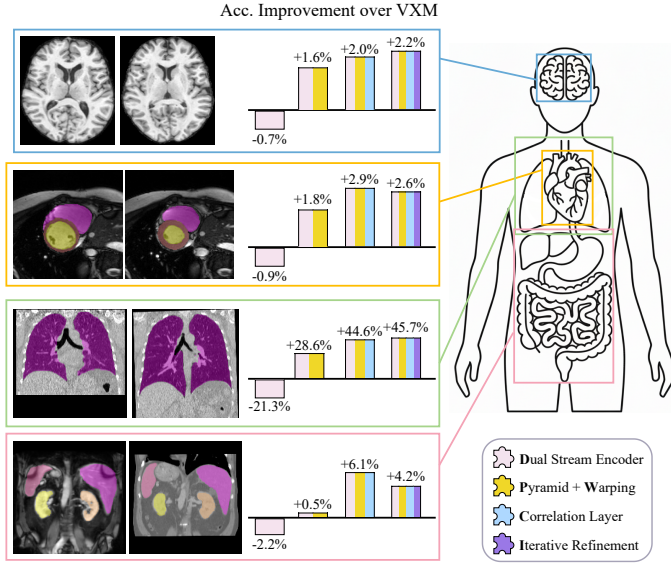


Fig. 1: Registration-specific designs substantially enhance registration performance (up to **45.7%**) over the VoxelMorph (VXM) baseline across a wide range of registration benchmarks covering multiple modalities and anatomies.

Medical image registration drives numerous clinical and research applications by estimating spatial correspondences between scans from different timepoints, subjects, modalities, or acquisitions (Jian et al., 2025; Avants et al., 2008; Pluim et al., 2003). Accurate registration enables quantitative analysis, disease monitoring, and treatment planning (Im et al., 2025; Baheti et al., 2021; Brock et al., 2017) across anatomies ranging from the brain and lungs to the heart and abdomen (Risholm et al., 2011; Reinhardt et al., 2008; Makela et al., 2002; Xu et al., 2016). With the emergence of deep learning, deformable registration has shifted from hand-crafted similarity metrics and iterative optimization to data-driven networks capable of learning spatial transformations directly from image pairs (Balakrishnan et al., 2019).

Despite impressive progress, a central question remains: what truly drives progress in learning-based medical image registration? Recent research has pursued mainly two paths:

The first emphasizes the adoption of “trend-driven” computational blocks: modules imported from the broader computer vision community, such as Vision Transformers (Dosovitskiy et al., 2020), large-kernel CNNs (Ding et al., 2022), Vision MLPs (Tolstikhin et al., 2021), and more recently, State Space Models (Mamba) (Gu and Dao, 2024). These blocks are introduced to capture long-range dependencies, expand receptive fields, mix spatial tokens efficiently, or perform linear-time sequence modeling. However, these general-purpose blocks were not developed specifically for the registration task.

The second revisits the incorporation of registration-specific designs, often well-established in “traditional”, optimization-based registration: principles developed decades ago to address the geometric nature of image alignment. These designs include dual-stream encoding to project images into a shared latent

space for matching, correlation layers to compute dense voxel-wise similarities, multi-resolution pyramids to handle large displacements, and iterative refinement to progressively improve deformation fields (Rueckert and Aljabar, 1999; Brox et al., 2004; Hosni et al., 2012).

In practice, most modern architectures combine both directions, resulting in intertwined contributions, raising the question: are performance gains driven primarily by novel computational blocks, or by classic, well-established registration-oriented strategies? A glimpse of two related domains offers compelling evidence. In medical image segmentation, nnU-Net demonstrated that well-optimized CNN-based U-Nets consistently outperform Transformer- or Mamba-based variants across diverse benchmarks (Isensee et al., 2021, 2024), underscoring the power of task-aligned design over architectural novelty. Likewise, in natural language processing, state-of-the-art large language models (LLMs) have advanced through scaling in parameter size and data size, while sharing the same domain-grounded core principle: attention mechanism in Transformers (Vaswani et al., 2017; Raschka, 2025). This parallel motivates our central hypothesis: progress in medical image registration may depend more on registration-oriented priors than on importing the latest architectural trends.

To investigate this hypothesis systematically, we develop a modular framework that disentangles and quantifies the respective contributions of “trend-driven” and *registration-specific* components. We evaluate methods across four representative registration tasks, each posing a unique challenge (Figure 1): (i) subtle anatomical alignment in brain MRI registration, (ii) large displacement, sliding-boundaries in inhale-exhale lung CT registration, (iii) dynamic spatio-temporal deformation in cardiac MRI registration, and (iv) large inter-organ shifts and intensity mismatches in abdominal MRI-CT registration. This task diversity, together with both in-domain and out-of-domain (zero-shot) evaluations, enables a holistic examination of model behavior across deformation scales, modalities, and clinical contexts. To ensure maximal fairness and transparency, all experiments follow a unified unsupervised training setup with identical preprocessing and loss functions. This extends our previous work (Jian et al., 2024), where weak supervision via Dice loss on brain labels was used. While such supervision can encourage anatomically faithful alignment within a dataset, Jena et al. (2024b); Liu et al. (2025) have shown that it does not translate to robustness under domain shift, as networks may overfit to dataset-specific label distribution. Furthermore, we expand beyond single-study evaluation by introducing an open, dynamic benchmark that hosts all models, configurations, training and evaluation processes. The benchmark enables plug-and-play integration of new methods and tasks, fostering reproducible, transparent, and fair comparisons across both in-domain and zero-shot generalization settings.

Our key contributions are threefold:

- **Modularized Disentanglement:** A modular framework that isolates and systematically evaluates the impact of low-level “trend-driven” computation blocks versus high-level *registration-specific* designs, enabling fine-grained performance attribution.

- **Cross-Domain Empirical Analysis:** Extensive experiments across brain, lung, cardiac, and abdomen, covering mono-modal and multi-modal settings, as well as both in-domain and out-of-domain (zero-shot) testing, reveal that domain-specific designs are the consistent drivers of accuracy, smoothness, and generalizability.
- **Open, Extensible Benchmark:** A public and extensible benchmark with all code and results released at <https://github.com/BailiangJ/rethink-reg>, establishing a foundation for fair, reproducible, and continuously evolving registration research.

2. Related Work

2.1. Trend-driven Blocks Imported from Computer Vision

Architectural innovations from computer vision have been increasingly adopted for deformable registration. These blocks typically aim to capture long-range dependencies, enlarge receptive fields, or improve sequence modeling efficiency, often by replacing the convolutional encoder in U-Net-style architectures.

2.1.1. Transformers and MLP-based Models

The Vision Transformer (ViT) (Dosovitskiy *et al.*, 2020) represents images as sequences of patches and models global context via self-attention. Early studies (Chen *et al.*, 2021; Mok and Chung, 2022) incorporated ViT for medical image registration, motivated by its ability to capture long-range dependencies and address large displacements. However, the quadratic computational cost of global attention limits the direct use of ViT in volumetric settings. To mitigate this, the Swin Transformer (Liu *et al.*, 2021), which computes attention within shifted, localized windows, has been widely adopted. Chen *et al.* (2022) proposed TransMorph, where the Swin Transformer encoder replaces the convolutional encoder, improving the volumetric registration capacity. Later extensions introduced cross-attention between source and target features or across multiple scales (Song *et al.*, 2022; Shi *et al.*, 2022; Chen *et al.*, 2023; Ghahremani *et al.*, 2024). Beyond attention-based models, Tolstikhin *et al.* (2021) proposed MLP-Mixer for spatial and channel mixing of image patches, reaching effective global context as well. Building on this, Meng *et al.* (2024) adapted MLPs to capture global dependencies in registration. While promising, these methods' architectural designs often incorporate registration-specific components like multi-resolution pyramids or correlation layers, making it difficult to isolate the precise contribution of the attention or MLP mechanism alone.

2.1.2. Convolutional Variants

VoxelMorph (Balakrishnan *et al.*, 2019) established the convolutional U-Net (Ronneberger *et al.*, 2015) as a widely used baseline for learning-based registration. Building on this, researchers have explored variations beyond the standard 3×3 convolutional operator. Miao *et al.* (2018) employed dilated convolutions (Yu and Koltun, 2015) to expand the receptive field without increasing parameters in registration networks,

while Jia *et al.* (2022) used large-kernel convolutions (Ding *et al.*, 2022) to approximate long-range dependencies similar to Transformer-based networks. Moreover, Huang *et al.* (2024) utilized deformable convolutions (Dai *et al.*, 2017) to allow adaptive kernel shapes for more flexible feature extraction during registration, while Cheng *et al.* (2025); Zheng *et al.* (2024) adjusted kernel weights to adapt to anatomical or modality-specific variations. While these approaches enhance the convolutional paradigm, they raise the question of how these incremental improvements to the backbone compare to the impact of higher-level architectural designs.

2.1.3. State-Space Models (Mamba)

Mamba (Gu and Dao, 2024) is a selective state-space model (SSM) that enables dynamic selectivity in sequential modeling. Its hardware-aware algorithm allows nearly linear complexity with respect to sequence length. Guo *et al.* (2024) introduced MambaMorph by substituting CNNs with Mamba blocks in the VoxelMorph's encoder, and subsequent works (Hu *et al.*, 2024; Wen *et al.*, 2024; Chen *et al.*, 2024, 2025b; Xu *et al.*, 2025; Xue *et al.*, 2025) explore Mamba for brain MRI, lung CT, and Abdomen CT registration. These studies posit that the efficient modeling of long-range dependencies by Mamba improves the registration results. However, the benefits usually come together with registration-specific designs like image pyramid, leaving the true benefit of Mamba for registration as an open question. A recent revisiting study on nnU-Net (Isensee *et al.*, 2024) reported that replacing convolution with Mamba layers (Ma *et al.*, 2024a) did not affect segmentation performance. The reported gains primarily stemmed from residual connections in the U-Net encoder.

2.1.4. Other Imported Paradigms

Beyond backbones, several other advanced techniques have been adapted for registration. Wolterink *et al.* (2022); Van Harten *et al.* (2023); van Harten *et al.* (2024); Sideri-Lampretsa *et al.* (2024) adopted Implicit Neural Representations (INRs) (Sitzmann *et al.*, 2020) to parametrize flow fields with coordinate-based MLPs, while Xu *et al.* (2021); Wu *et al.* (2022) add NeuralODE (Chen *et al.*, 2018) on top to integrate the velocity field and achieve smoother and more accurate deformations. Besides, Kim *et al.* (2022); Qin and Li (2023) exploit the latent features learned in the Diffusion-based generative models to further refine the registration. Together, these methods illustrate the breadth of paradigms imported from computer vision into registration, even though their specific contributions are often intertwined with domain-specific adaptations.

2.2. Registration-specific Architectural Design

In parallel to trend-driven imports, registration-specific designs explicitly encode domain knowledge about the image alignment task itself. These strategies build primarily on principles established in classical registration and optical flow estimation methods, and have been widely inherited and extended in deep learning frameworks.

2.2.1. Dual-stream/Siamese Encoders

Instead of concatenating the source and target images at the input, dual-stream encoders process them independently with shared or separate weights. This design is common in optical flow networks (Sun et al., 2018; Teed and Deng, 2020) and has been adapted in registration frameworks such as Pan et al. (2021); Meng et al. (2022); Shi et al. (2022); Pan et al. (2024). It encourages the extraction of feature representations that remain discriminative within each image but comparable across images, thereby facilitating explicit correspondence matching, as in traditional feature-based image matching methods (Lowe, 2004; Bay et al., 2006).

2.2.2. Motion Pyramid and Warping

Coarse-to-fine refinement is a classical paradigm in registration (Rueckert and Aljabar, 1999; Avants et al., 2008; Vercauteren et al., 2009). Deformable image registration is inherently a high-dimensional, nonlinear, and non-convex optimization problem, often leading to ill-conditioning and susceptibility to local minima. Multi-resolution optimization mitigates these challenges by gradually increasing spatial resolution, analogous to simulated annealing or scale-space optimization, which improves convergence to better optima (Jena et al., 2024a). Deep learning counterparts adopt hierarchical pyramids of images or features: deformations are first estimated at coarse resolution to capture global motion, and then progressively refined at finer scales using warped features or images from the previous stage (Mok and Chung, 2020b; Kang et al., 2022; Lv et al., 2022; Meng et al., 2022; Yin et al., 2023). This design is particularly effective for capturing large displacements.

2.2.3. Correlation Layers / Cost Volumes

Correlation volumes, long established in optical flow estimation (Brox et al., 2004; Hosni et al., 2012; Dosovitskiy et al., 2015; Sun et al., 2018; Teed and Deng, 2020), have been adapted for solving voxel-wise correspondence in registration. Heinrich (2019) approximated cost volumes via localized MSE, while later works (Wang et al., 2023b; Liu et al., 2024b; Jian et al., 2024) directly compute inner products between features within search windows. Correlation is often paired with dual-stream encoders, enabling explicit voxel-wise comparison between target and source features rather than relying on implicit fusion.

2.2.4. Iterative/Recursive Refinement

Complementing the multi-scale approach, iterative refinement improves alignment by applying recursive updates at a fixed resolution. Inspired by recurrent updates in optical flow (Teed and Deng, 2020), this strategy iteratively refines the deformation by repeatedly feeding the currently warped image or features back into a refinement network or a decoder block to predict incremental updates (Zhao et al., 2019; Jia et al., 2021; Ma et al., 2023; Wang et al., 2024; Ma et al., 2024b). This process progressively corrects residual misalignment at the same scale, and conceptually approximates velocity integration techniques such as scaling-and-squaring, often yielding smoother

and more accurate deformations than a single, one-shot prediction.

2.2.5. Other Task-specific Strategies

Other strategies aim to enforce deformation plausibility and consistency. Symmetric registration (Avants et al., 2008; Mok and Chung, 2020a) jointly deforms both images to a mid-space, while inverse consistency can be enforced either by construction (Greer et al., 2023; Honkamaa and Marttinen, 2023) or by extra constraints (Kim et al., 2021; Greer et al., 2021; Van Harten et al., 2023; Tian et al., 2023; Reithmeir et al., 2025). Parametric formulations based on splines (Qiu et al., 2025) or Gaussian primitives (Li et al., 2025; Tian et al., 2025) further provide smooth, compact, and efficient parameterizations of the deformation field. Collectively, these strategies illustrate the diversity of task-specific solutions developed to improve alignment quality and regularity.

2.3. Registration Benchmarks

Public registration challenges have played an essential role in advancing the field by providing standardized datasets, tasks, and evaluation protocols across anatomies and modalities. Notable examples, including EMPIRE10 (Murphy et al., 2011), Learn2Reg (Hering et al., 2022; Hansen et al., 2025), REMIND2Reg (Dorent et al., 2025), and LUMIR (Chen et al., 2025a) cover anatomies including brain, lung, chest, and abdomen, and modalities including MRI, CT, and Ultrasound. These benchmarks offer a clear snapshot of the state-of-the-art on difficult clinical problems and have fostered rapid methodological innovation. At the same time, the primary goal of challenge participation is to maximize leaderboard performance. To achieve this, competing solutions often integrate a range of architectural choices, supervision schemes, and training strategies into complex pipelines. While effective, this makes it difficult to attribute performance gains to specific design components, obscuring the individual contributions of trend-driven modules or registration-specific elements. This observation motivates our study: rather than evaluating holistic solutions, we propose a modular framework that isolates and quantifies the effects of different architectural components under identical experimental setups. In doing so, we aim to provide a clearer understanding of which design choices truly drive performance in deformable registration.

3. Methodology

3.1. Preliminaries

Let $I_t, I_s : \Omega \rightarrow \mathbb{R}$ be the target (or fixed) and source (or moving) images, respectively, defined over a spatial domain $\Omega \subset \mathbb{R}^d$, where the dimension $d \in \mathbb{N}$ is typically 2 or 3. Deformable image registration aims to find a dense deformation field $\phi : \Omega \rightarrow \Omega$ that establishes a spatial correspondence between the two images.

In a deep learning-based framework, this is achieved by training a neural network Φ , parametrized by weights θ , to predict the deformation field from the input image pair:

$$\phi = \Phi_{\theta}(I_t, I_s). \quad (1)$$

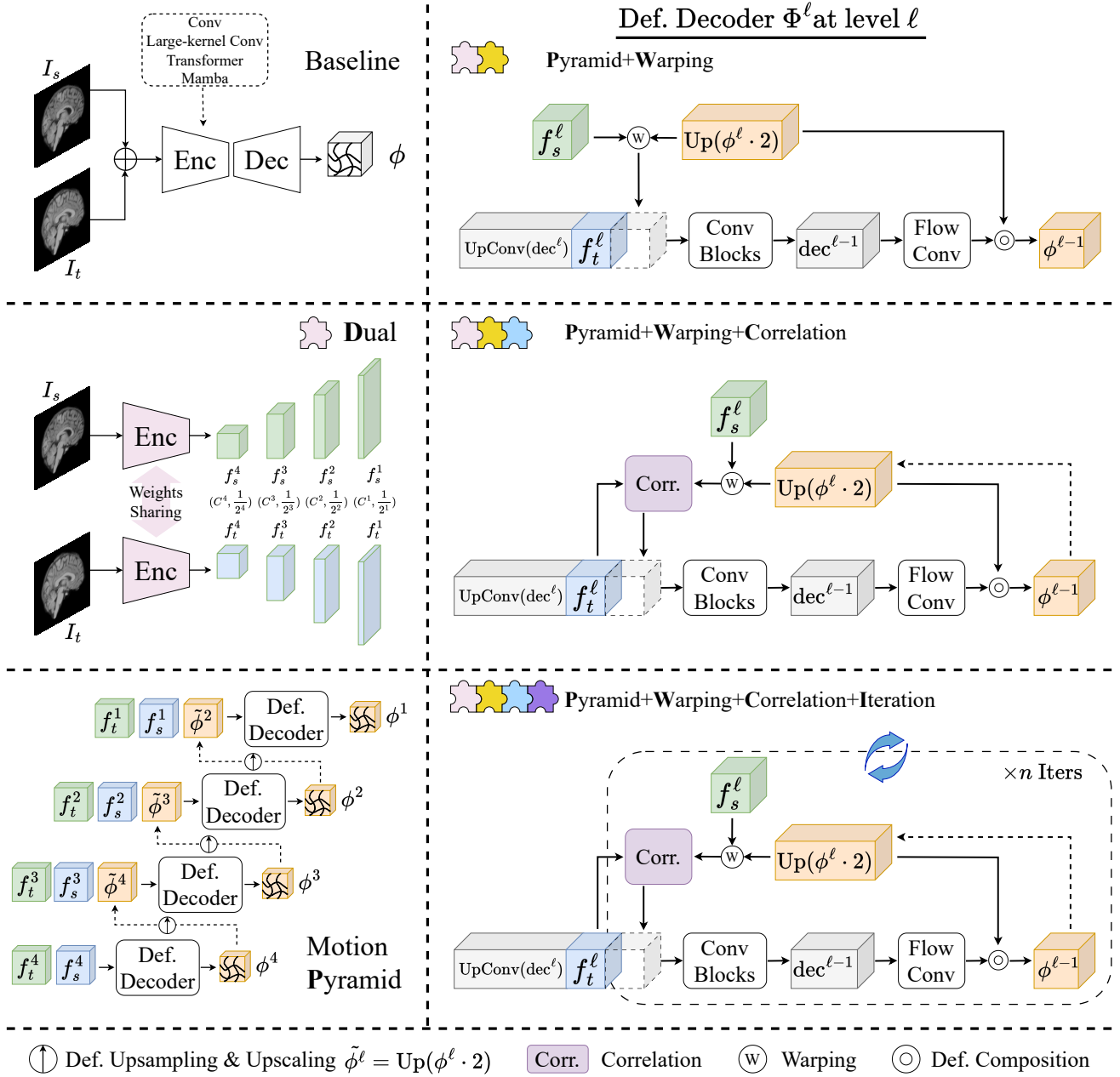


Fig. 2: Overview of the baseline and the modular registration components. **Upper Left:** the *Baseline* architecture concatenates image pairs and predicts the deformation field directly. **Middle Left:** *Dual*-stream encoder extracts multi-resolution feature pyramids for the source and target images separately. **Bottom Left:** *Motion Pyramid* refines the deformation field progressively from coarse to fine using hierarchical features. **Right:** workflow of the deformation decoder at level ℓ . Level ℓ corresponds to $2^{-\ell}$ resolution. Registration-specific designs are cumulatively integrated: **Top:** *Pyramid and Warping*; **Middle:** *Correlation*; **Bottom:** *Iteration*. The cube widths are illustrative and do not represent the exact channel dimensions.

In this work, we adopt the common backward (or pullback or Euler coordinates) convention: ϕ maps each coordinate $\mathbf{x} \in \Omega$ in the target image's space to its corresponding coordinate in the source image's space. The source image is then warped according to this field to align with the target. The warped source (or moved) image I_m is defined as:

$$I_m(\mathbf{x}) = (I_s \circ \phi)(\mathbf{x}) := I_s(\phi(\mathbf{x})). \quad (2)$$

Equivalently, the deformation field can be parametrized via an

additive displacement field $\mathbf{u} : \Omega \rightarrow \mathbb{R}^d$, $\phi(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$. The objective is to find a ϕ such that the moved image I_m is similar to the target I_t , i.e., $I_t \approx I_s \circ \phi$.

To optimize the network parameters θ , a loss function $\mathcal{L}$ is minimized via gradient descent. This loss typically consists of two main components:

$$\mathcal{L}(\phi, I_t, I_s) = \mathcal{L}_{\text{sim}}(I_t, I_s \circ \phi) + \lambda \mathcal{L}_{\text{reg}}(\phi). \quad (3)$$

The first term, $\mathcal{L}_{\text{sim}}$, quantifies the dissimilarity between the

target image and the warped source (moved) image. Its choice depends on the imaging modalities; common examples include Mean Squared Error (MSE) or Localized Normalized Cross-Correlation (LNCC) for mono-modal registration, and Mutual Information (MI) or MIND-SSC for multi-modal cases.

The second term, $\mathcal{L}_{\text{reg}}$, is a regularization penalty imposed on the deformation field ϕ or the displacement field $\mathbf{u}$. This term ensures the transformation is smooth and physically plausible, penalizes unrealistic distortions such as tears or foldings. Common regularizers are Diffusion (first-order smoothness), Bending energy (second-order smoothness), and a direct penalty on non-positive Jacobian determinants. The hyperparameter λ controls the trade-off between image similarity and deformation regularity.

While additional supervision can be incorporated using annotations like segmentation labels, this work focuses on unsupervised training, where the optimization is driven purely by information from the image intensities.

3.2. Baseline

Our reference model is the canonical U-Net-based architecture introduced by VoxelMorph (Balakrishnan *et al.*, 2019), a fully-convolutional encoder-decoder network with skip connections. Conceptually, this architecture operates as a single-stream model where the target and source images, I_t and I_s , are concatenated and fed into the network as a single tensor. The encoder progressively extracts hierarchical features using a series of standard 3×3 convolutional blocks, and the decoder up-samples and refines these features to predict a final voxel-wise displacement field $\mathbf{u}$. This displacement defines the dense deformation field via $\phi(\mathbf{x}) = \mathbf{x} + \mathbf{u}(\mathbf{x})$, which is then used to warp the source image.

This simple yet powerful CNN-based U-Net serves as our baseline framework. It enables systematic integration of both trend-driven computational blocks (e.g., Vision Transformers, Mamba) and registration-specific designs (e.g., dual-stream image encoder, motion pyramid, correlation layers). All subsequent variants are derived by modular substitution or extension of this baseline.

3.3. Modular Framework

To disentangle the contributions of computational novelty versus registration-specific design, we adopt a modular evaluation strategy. Each candidate block is integrated into the baseline U-Net in isolation or in controlled combinations, while all other components (loss functions, training datasets, hyperparameters, and random seeds) are held fixed.

Formally, let $\mathcal{M}_{\text{base}}$ denote the baseline U-Net registration model (VoxelMorph). A trend-driven block $\mathcal{B}_{\text{trend}}$ or a registration-specific block $\mathcal{B}_{\text{reg}}$ can be substituted into the baseline to yield a new variant:

$$\mathcal{M} = \mathcal{M}_{\text{base}} \oplus \mathcal{B}_1 \oplus \mathcal{B}_2 \oplus \cdots \oplus \mathcal{B}_n,$$

where each $\mathcal{B}_i \in \{\mathcal{B}_{\text{trend}}, \mathcal{B}_{\text{reg}}\}$ represents a specific architectural component, and $\oplus$ denotes modular replacement or augmentation. By progressively stacking blocks from either category, we can measure both marginal contributions (the effect of

each block individually) and interaction effects (when multiple blocks are combined).

This modular logic yields a factorial comparison design, enabling fair and transparent attribution of performance gains to specific architectural choices. Importantly, every variant is trained and evaluated under identical conditions, ensuring that observed differences reflect only the architectural modifications.

3.4. Trend-driven Computational Blocks

Recent advances in computer vision have introduced a series of architectural building blocks designed to expand receptive fields, capture long-range dependencies, or improve sequence modeling efficiency. To assess their impact on deformable registration, we replace the encoder blocks of the baseline U-Net (VoxelMorph) with the following variants.

Vision Transformer (TransMorph). To capture global dependencies, TransMorph (Chen *et al.*, 2022) replaces the convolutional encoder with Swin Transformer blocks. The Swin Transformer divides images into non-overlapping patches, similar to the standard Vision Transformer, but restricts self-attention to localized, shifted windows. This design efficiently models both local and global context, while avoiding the quadratic complexity of full attention. The decoder remains convolutional for comparability. This variant is denoted as **TM**.

Large-Kernel Convolution (LKU-Net). Larger convolutional kernels can approximate long-range dependencies (Ding *et al.*, 2022). LKU-Net (Jia *et al.*, 2022) augments each encoder block with a parallel 5×5 convolution branch, effectively enlarging the receptive field while retaining the efficiency of convolutions. This variant is denoted as **LKU**.

Mamba Encoder. Mamba (Gu and Dao, 2024) is a state-space model (SSM) with a hardware-optimized selective scanning mechanism for efficient sequence modeling. MambaMorph (Guo *et al.*, 2024) first introduced it for deformable registration to handle long-range dependencies more efficiently. We evaluate two integration strategies:

- **Mam-VXM:** Replaces the convolutional encoder of VoxelMorph with stacked Mamba blocks, while retaining the original decoder.
- **Mam-TM:** Replaces the Swin Transformer blocks in TransMorph with Mamba blocks. Patch embedding, positional encodings, and MLP layers are retained to preserve the patch-based formulation.

These two settings allow us to disentangle the role of Mamba within convolutional and transformer-style encoders.

3.5. Registration-specific Design Blocks

Beyond generic trend-driven blocks, several architectural strategies have been proposed specifically for deformable registration. These strategies encode domain knowledge, such as coarse-to-fine refinement, explicit feature correspondences via correlation, or iterative updates, and are well-established in

“traditional” registration. We integrate the following four components into our modular framework:

To formalize composition, let $\phi_1, \phi_2 : \Omega \rightarrow \Omega$ be deformation fields. Their composition is

$$(\phi_1 \circ \phi_2)(\mathbf{x}) = \phi_2(\phi_1(\mathbf{x})), \quad (4)$$

corresponding to applying ϕ_1 first and then ϕ_2 . In terms of displacement fields $\mathbf{u}_1, \mathbf{u}_2$, this becomes

$$(\mathbf{u}_1 \circ \mathbf{u}_2)(\mathbf{x}) = \mathbf{u}_1(\mathbf{x}) + \mathbf{u}_2(\mathbf{x} + \mathbf{u}_1(\mathbf{x})). \quad (5)$$

Dual-Stream Encoders. Instead of concatenating the target and source images, a shared-weight encoder E extracts a multi-scale feature pyramid from the source and target images independently:

$$\{f_t^\ell\}_{\ell=0}^L = E(I_t), \quad \{f_s^\ell\}_{\ell=0}^L = E(I_s). \quad (6)$$

where $f_t^\ell, f_s^\ell \in \mathbb{R}^{H_\ell \times W_\ell \times D_\ell \times C_\ell}$ denotes the feature map at resolution level ℓ , with spatial size (H_ℓ, W_ℓ, D_ℓ) and channel dimension C_ℓ , where D_ℓ is present only in 3D cases. Level ℓ corresponds to scale $2^{-\ell}$ resolution relative to the original image. This independent encoding encourages discriminative yet comparable features for subsequent matching. Denoted as **D**.

Motion Pyramid and Warping. To capture large deformations, we adopt a coarse-to-fine motion pyramid. At the coarsest level $\ell = L$, the decoder predicts ϕ^L . For each finer level ℓ , the flow from level $\ell + 1$ is upsampled and upscaled:

$$\tilde{\phi}^{\ell+1} = 2 \cdot \text{Up}(\phi^{\ell+1}). \quad (7)$$

The decoder Φ^ℓ then predicts a residual update

$$\Delta\phi^\ell = \Phi^\ell(f_t^\ell, f_s^\ell \circ \tilde{\phi}^{\ell+1}), \quad (8)$$

and the flow is refined as

$$\phi^\ell = \tilde{\phi}^{\ell+1} \circ \Delta\phi^\ell. \quad (9)$$

This component is denoted as **PW**.

Correlation Layers. To explicitly establish voxel-wise correspondences, we compute local correlations between features at level ℓ . Let

$$f_t^\ell, f_s^\ell \in \mathbb{R}^{H_\ell \times W_\ell \times D_\ell \times C_\ell} \quad (10)$$

be the target and source feature maps. For each voxel $\mathbf{x} \in \Omega_\ell$, the correlation is defined within a local neighborhood $\mathcal{N}_r(\mathbf{x}) = \{\mathbf{y} \in \Omega_\ell : \|\mathbf{y} - \mathbf{x}\|_\infty \leq r\}$ of radius r as

$$C^\ell(\mathbf{x}, \mathbf{y}) = \frac{1}{C_\ell} f_t^\ell(\mathbf{x})^\top f_s^\ell(\mathbf{y}), \quad \mathbf{y} \in \mathcal{N}_r(\mathbf{x}). \quad (11)$$

The resulting correlation volume has shape $H_\ell \times W_\ell \times D_\ell \times (2r+1)^d$, where $d \in \{2, 3\}$ is the spatial dimension. This module is denoted as **C**.

Iterative Refinement. Instead of estimating the full deformation field in a single forward pass, we employ iterative refinement at each level. Let Φ^ℓ denote the decoder at level ℓ . At

iteration k , given the current estimate $\phi_{(k)}^\ell$, Φ^ℓ predicts an incremental update

$$\Delta\phi_{(k)}^\ell = \Phi^\ell(f_t^\ell, f_s^\ell \circ \phi_{(k)}^\ell), \quad (12)$$

and the field is updated by composition

$$\phi_{(k+1)}^\ell = \phi_{(k)}^\ell \circ \Delta\phi_{(k)}^\ell. \quad (13)$$

Starting from $\phi_{(0)}^\ell = \tilde{\phi}^{\ell+1} = \text{Up}(\phi^{\ell+1}) \cdot 2$, this recursive refinement progressively improves alignment. Conceptually, the procedure approximates the integration of an underlying velocity field, yielding smoother deformations. Denoted as **I**.

Modular Integration. Each of the above registration-specific components (**D**, **PW**, **C**, **I**) can be integrated into the baseline independently or in combination. As shown in Figure 2, within our framework, these modules are incrementally added on top of the baseline encoder-decoder, enabling systematic quantification of their individual contributions as well as their synergistic effects.

4. Experiments

4.1. Datasets and Tasks

We evaluate across four representative deformable registration tasks covering diverse anatomies, modalities, and motion patterns: (1) cross-sectional brain MRI registration, (2) longitudinal lung CT registration, (3) temporal cardiac MRI 2D registration, and (4) intra-subject abdominal MRI–CT registration. This diversity allows us to assess both trend-driven and registration-specific designs under heterogeneous imaging and anatomical conditions.

4.1.1. Cross-sectional Brain MRI Registration

This task targets mono-modal brain MRI alignment across different subjects. We use the LUMIR challenge dataset [Chen et al. \(2025a\)](#) as the primary training data, taking the first 500 scans for training from 3384 subjects. Evaluation is conducted on the 30 leaderboard validation subjects, which include 10 in-domain, 10 out-of-domain with landmark annotations, and 10 out-of-domain high-field MRI scans with manual segmentations, resulting in 27 validation pairs. To test generalization, we further perform zero-shot evaluation on OASIS [Marcus et al. \(2007\)](#) (84 scans), ADNI [Jack Jr et al. \(2008\)](#) (43), IXI¹ (115), LPBA [Shattuck et al. \(2008\)](#) (40), and Mindboggle [Klein et al. \(2005\)](#) (100) datasets, each with 200 randomly sampled test pairs following [Jian et al. \(2024\)](#). All scans are in 1 mm isotropic spacing and with image size $160 \times 192 \times 224$. Intensities are min–max normalized between the 0th and 99.99th percentiles without clipping. Training uses the Local Normalized Cross-Correlation (LNCC) loss with window size 9 and a regularization weight $\lambda = 0.5$. The batch size is set to 1, with 100 epochs of 250 iterations each. Left–right flipping ($p = 0.5$) is applied as data augmentation.

¹<https://brain-development.org/ixi-dataset/>

4.1.2. Longitudinal Lung CT Registration

We evaluate intra-patient registration between inspiration and expiration phases using the NLST dataset from the Learn2Reg challenge (Clark *et al.* (2013); Hering *et al.* (2022)). The dataset provides paired CT scans, lung masks, and anatomical keypoints. We split the data into 110/50/50 pairs for training, validation, and testing. For zero-shot evaluation, we use the Lung250M-4B dataset (Falta *et al.* (2023)), consisting of 17 pairs from the validation and test sets (excluding overlap with the NLST dataset) with landmark annotations. All images in the NLST dataset are in 1.5 mm isotropic spacing and with image size $224 \times 192 \times 224$. Because voxel spacing in Lung250M-4B varies, we resize the images to an identical image size $224 \times 192 \times 224$. Therefore, we report Target Registration Error (TRE) in millimeters for the NLST dataset and in voxel units for the resampled Lung250M-4B dataset. Intensities are clipped to the Hounsfield Unit (HU) range of $[-1000, 500]$ and normalized to $[0, 1]$. The LNCC loss (window size = 9) is used with a regularization weight $\lambda = 1.0$. Training uses a batch size of 1 for 200 epochs (80 iterations per epoch) and left-right flipping ($p = 0.5$) augmentation.

4.1.3. Temporal Cardiac MRI 2D Registration

This task evaluates short-axis cardiac motion between end-diastole (ED) and end-systole (ES) frames. We use the ACDC dataset (Bernard *et al.* (2018)) for training and in-domain testing and the M&Ms (Multi-Centre, Multi-Vendor & Multi-Disease) dataset (Campello *et al.* (2021)) for zero-shot testing. Both datasets include segmentations of the left ventricle (LV), right ventricle (RV), and myocardium (MYO). The ACDC dataset is split into 80/20/50 subjects for training, validation, and testing, corresponding to 506/190/194 slices, while the M&Ms dataset contains 1859 slices from six centers across three countries. Because of thick-slice acquisitions (5–10 mm) along the cardiac long-axis, registration is performed in 2D slices. Slices are cropped to the cardiac region of interest based on segmentation masks, resulting in an image size 128×128 . Intensities are min-max normalized to $[0, 1]$ per slice. The Mean Squared Error (MSE) loss is used with a regularization weight $\lambda = 0.05$. Each iteration processes all 2D slices from one 3D volume as a batch. Training runs for 200 epochs with 80 iterations per epoch, with random horizontal and vertical flipping for augmentation.

4.1.4. Intra-subject Abdomen MRI-CT Registration

We use the abdominal MRI-CT pairs from the Learn2Reg challenge (Hering *et al.* (2022)), comprising 16 intra-patient cases. Eight pairs include manual annotations of the spleen, liver, and kidneys, which serve as the test set. The remaining eight pairs are segmented using TotalSegmentator (Wasserthal *et al.* (2023)), and used for training. To remove coarse misalignment and focus on deformable correspondence, we perform an initial rigid translation alignment using organ centroid coordinates. All volumes are in 2 mm isotropic spacing and with image size $192 \times 160 \times 192$. MRI intensities are min-max normalized between the 0th and 99.99th percentiles, while CT intensities are clipped to the HU range $[-450, 450]$ and normalized

to $[0, 1]$. The Modality Independent Neighborhood Descriptor (MIND) loss (Heinrich *et al.* (2013)) is used with radius = 2, dilation = 2, and a regularization rate $\lambda = 1.0$. Training uses batch size 1, 200 epochs (8 iterations per epoch), and left-right flipping ($p = 0.5$) as augmentation.

4.2. Modular Framework and Architectural Configurations

To ensure a fair comparison, all tested architectures adhere to two universal design principles. First, all models predict final displacement fields at half-resolution, which are then upsampled to the full image grid for warping. Second, all models share a standardized output layer, the *Flow Convolution*: a kernel size 3 convolution with stride 1, output channels equal to the spatial dimension, weight initialization $\mathcal{N}(0, 10^{-5})$, and zero bias.

4.2.1. Low-level Computational Blocks

The baseline model is **VoxelMorph (VXM)** (Balakrishnan *et al.* (2019)), implemented as a vanilla U-Net (Ronneberger *et al.* (2015)) with convolutional downsampling and transposed-convolution upsampling layers following Jian *et al.* (2024). The encoder and decoder channel numbers are set to $[16, 32, 64, 96, 128]$ and $[128, 96, 64, 32]$, respectively, followed by two remaining convolutional layers (32 channels) before the *Flow Convolution* layer that predicts the displacement field.

We then replace the convolutional encoder of the baseline with various computational backbones while keeping all other components identical: - **Mam-VXM**: substitutes the CNN encoder with *Mamba* (Gu and Dao (2024)) blocks, while preserving the VXM decoder. - **TransMorph (TM)** (Chen *et al.* (2022)): replaces the convolutional encoder with a Swin Transformer encoder. The original implementation² (embedding dimension 96) is used. - **Mam-TM**: identical to TransMorph except that Swin Transformer blocks are replaced with Mamba blocks. - **LKU-Net** (Jia *et al.* (2022)): adds parallel large-kernel convolutions ($k = 5$) to enlarge the receptive field. The original implementation³ (initial feature channel 16) is used.

4.2.2. High-level Registration-specific Designs

We build all variants on top of the canonical VXM backbone to isolate the effect of each registration-specific design component. The following modules are incrementally added:

Dual-stream Encoder (D). Instead of concatenating source and target images at the input, two parallel encoders with halved feature channels as VXM and shared weights, process them independently. The resulting feature pyramids are concatenated at each level before decoding. The pyramids consist of features at resolutions $[1/16, 1/8, 1/4, 1/2]$.

²https://github.com/junyuchen245/TransMorph_Transformer_for_Medical_Image_Registration

³<https://github.com/xi-jia/LKU-Net>

Motion Pyramid and Warping (P). A hierarchical motion estimation scheme is introduced, producing deformation fields at multiple resolutions (1/16, 1/8, 1/4, 1/2). The remaining 2 convolution layers are removed, and the *Flow Convolution* is added at each level ℓ to predict the multi-scale flow fields ϕ^ℓ . The low-resolution flow is upsampled and used to warp the moving feature map before the next refinement stage.

Correlation Volume (C). A localized correlation layer computes the inner product between source and target features within a local search window, forming a correlation volume. The target features serve as the context and are concatenated with the correlation volume, which together are passed to the decoder to enhance voxel-wise feature correspondence without explicit supervision. The search window of the correlation layer at each resolution level is set using a radius of $r = 1$ for simplicity and efficiency.

Iterative Refinement (I). Iterative refinement applies recursive updates to the deformation field at a fixed resolution. The currently warped features are fed back into the decoder for an additional refinement pass. We set $n = 2$ iterations at 1/4 and 1/2 resolutions following Jian *et al.* (2024).

We intentionally use a consistent set of hyperparameters across all model variants rather than performing exhaustive, per-model tuning. The goal of this study is a *ceteris paribus* (all other things being equal) comparison to isolate the impact of architectural choices, not to find the peak performance of each individual configuration. This ensures that any observed performance differences are attributable to the components being tested.

4.3. Training Protocol

All models are trained in a fully unsupervised manner using the training loss $\mathcal{L} = \mathcal{L}_{\text{sim}} + \lambda \mathcal{L}_{\text{reg}}$. We use the diffusion regularizer on the spatial gradients of the displacement $\mathcal{L}_{\text{reg}} = \|\nabla \mathbf{u}\|^2$ as the smoothness regularization, and select $\mathcal{L}_{\text{sim}}$ and λ based on the modality and task.

Pyramidal Training Loss. For architectures containing motion pyramids, we compute losses at each resolution level ℓ and weight them proportionally to the scale:

$$\mathcal{L}_{\text{pyramid}} = \sum_{\ell} 2^{-\ell} \left(\mathcal{L}_{\text{sim}}(I_t^\ell, I_s^\ell \circ \phi^\ell) + \lambda \mathcal{L}_{\text{reg}}(\phi^\ell) \right), \quad (14)$$

where downsampled images I_t^ℓ and I_s^ℓ are obtained by average pooling.

Optimization and Fairness. All experiments use the Adam optimizer with an initial learning rate of 10^{-4} and an exponential decay rate of 0.996. The random seed for both training and testing is fixed at 2023, ensuring that each method is trained on exactly the same randomly sampled image pairs per iteration. A bidirectional training strategy is adopted. Each training iteration performs both forward ($I_s \rightarrow I_t$) and backward ($I_t \rightarrow I_s$) registrations, with the total loss averaged across directions. This setting guarantees a fair, controlled comparison among all model variants.

Table 1: Architectural profiles of the evaluated methods. We report the number of trainable parameters (Params, in millions), floating-point operations (FLOPs, in GB), peak GPU memory usage during training (Max. Mem Train, in GB), and inference runtime per image pair (Runtime Infer, in seconds). All values are measured on images of size $160 \times 192 \times 224$.

	Params (M)	FLOPs (GB)	Max.Mem Train (GB)	Runtime Infer (s)
VXM	2.64	470	6.23	0.17
Mam-VXM	2.38	462	8.56	0.09
TM	46.56	612	10.24	0.20
Mam-TM	46.77	590	8.02	0.06
LKU	2.07	263	5.12	0.13
Dual	1.92	433	5.13	0.28
DWP	1.89	339	5.07	0.25
DWCP	1.71	352	5.35	0.31
DWCPI	1.71	646	5.35	0.50

4.4. Evaluation Protocol

We assess each model in terms of registration accuracy, deformation plausibility, and computational efficiency (Rohlfing (2011)).

Registration Accuracy. We compute: (1) Dice similarity coefficient (DSC, %), (2) 95th percentile Hausdorff distance (HD95), (3) average symmetric surface distance (ASSD), (4) normalized surface dice (NSD), and (5) target registration error (TRE) for keypoint annotations. Labels and keypoints are used exclusively for evaluation and not during training.

Deformation Smoothness. We quantify the regularity of the deformation fields using: (1) the standard deviation of the log-Jacobian determinant (SDlogJ) (Leow *et al.* (2007)), and (2) the ratio of non-diffeomorphic voxels (NDV) (Liu *et al.* (2024a)). Lower values indicate smoother and more physically plausible transformations.

Qualitative Visualization. For qualitative comparison, we visualize: (1) target and warped source contours, (2) intensity difference maps for mono-modal registration, and (3) deformation grids. This allows visual assessment of alignment quality and the smoothness of local displacements.

Model Efficiency Profiling. To decouple accuracy gains from increased computational cost, we report for each method: (1) number of trainable parameters (Params, M), (2) number of floating-point operations (FLOPs, GB), (3) maximum GPU memory during training and inference (GB)

4.5. Implementation Details

All experiments are implemented in PyTorch 2.2 with CUDA 12.1 and trained using mixed precision. Training and inference are performed on NVIDIA A100 GPUs (40 GB memory). Each experiment is executed under the same environment, and all random seeds and preprocessing parameters are fixed to ensure full reproducibility.

We release all model configurations, training scripts, and evaluation code at <https://github.com/BailiangJ/rethink-reg> to facilitate transparent benchmarking.

5. Results

5.1. Cross-sectional Brain MRI Registration

Table 2 and Table 3 summarize the quantitative registration performance across six independent brain MRI datasets. Consistent with our previous findings in Jian *et al.* (2024), even under an unsupervised training setup, trend-driven computational blocks, such as Mamba and large-kernel convolution, yield only marginal accuracy gains, or even slight degradation, over the VXM baseline. In contrast, high-level registration-specific designs yield substantially stronger and more consistent improvements across all datasets.

Among these designs, the dual-stream encoder (**D**) yields a modest Dice decrease, possibly due to its halved feature capacity. Introducing the motion pyramid and warping (**WP**) markedly boosts accuracy (e.g., +3-4% DSC on LPBA) via coarse-to-fine refinement. Correlation volumes (**C**) further strengthen voxel-wise correspondence (+0.5-1% DSC) and improve deformation plausibility, reflected by reduced SDlogJ and NDV.

Iterative refinement (**I**) provides a final nudge (e.g., +0.1-0.3% DSC on OASIS and ADNI), indicating that recursive updates help capture subtle anatomical variations. However, this refinement occasionally increases SDlogJ, as observed on OASIS, suggesting slightly less consistent local smoothness despite fewer foldings (NDV $\approx$ 0).

The qualitative visualizations in Figure 3 further corroborate these trends: incorporating registration-specific components cumulatively improves spatial alignment and yields more anatomically plausible deformations.

5.2. Longitudinal Lung CT Registration

Table 4 summarizes the results for intra-subject lung registration on the NLST test set and the zero-shot Lung250M-4B dataset. Overall, the results exhibit consistent trends across both in-domain and out-of-distribution (OOD) evaluations. Compared to the baseline CNN-based VoxelMorph (VXM), Transformer- and Mamba-based models achieve marginally improved alignment accuracy on the NLST test set, suggesting that their ability to capture long-range dependencies can better accommodate large respiratory motions. However, these gains are modest and come at a substantially higher computational cost. As illustrated in Figure 5, registration-specific designs (**DWP**, **DWCP**, **DWCPI**) consistently achieve markedly lower 75th percentile in the cumulative TRE distribution, indicating more accurate alignment for the majority of anatomical keypoints.

In contrast, registration-specific architectures yield more pronounced and systematic improvements. Incorporating the motion pyramid and warping modules (**WP**) reduces the mean TRE from 3.6 mm to 2.6 mm on NLST, accompanied by a

+1% increase in DSC, demonstrating that the coarse-to-fine refinement efficiently captures large deformations. Adding a correlation volume (**C**) further strengthens voxel-level correspondence, lowering the TRE to approximately 2.0 mm while improving deformation regularity. Notably, this design also generalizes best to the unseen Lung250M-4B dataset, where the **DWCP** variant achieves the lowest TRE and most plausible deformations, confirming the robustness of local correlation-based matching under domain shift.

The iterative refinement module (**I**) provides marginal gains on the in-domain NLST set but yields slightly degraded keypoint accuracy in the OOD scenario. Interestingly, although **DWCPI** exhibits slightly higher TRE than **DWCP** on Lung250M-4B, qualitative inspection in Figure 4 reveals the best alignment in the left lung. This discrepancy might suggest that the iterative refinement stage encourages deformations that better preserve overall shape and boundary continuity, while relaxing on the localized keypoint correspondences. A similar effect is observed in LKU-Net, which gets weaker TRE results but better overall lung alignment. This is likely due to its enlarged receptive field that promotes smooth, large-scale alignment at the cost of reduced local precision.

In summary, while long-range models (Transformers, Mamba) modestly improve within-domain performance, the most reliable accuracy-regularity trade-off and generalization are consistently achieved by the registration-specific priors, particularly the pyramidal and correlation-based designs.

5.3. Temporal Cardiac MRI 2D Registration

Table 5 summarize results for the 2D cardiac MRI registration. The ACDC dataset represents the in-domain evaluation, while M&Ms serves as a challenging zero-shot generalization test involving data from different scanners, countries, and patient pathologies. The task demands accurate alignment between end-diastolic (ED) and end-systolic (ES) phases, which involves complex non-rigid motions of the left ventricle (LV), right ventricle (RV), and myocardium.

Compared to convolutional baselines, Transformer- and Mamba-based models achieve notable gains in both Dice and NSD metrics (+1.5-3% DSC and +0.5-1% NSD). These improvements reflect their capability to capture long-range dependencies that better model the global contraction-expansion dynamics of cardiac motion. However, LKU-Net with large kernel convolutions exhibits slightly lower accuracy despite smoother deformation fields, suggesting that the enlarged receptive field may come at the cost of reduced sensitivity to local motion details critical for anatomical alignment.

Nonetheless, these trend-driven improvements are ultimately matched and surpassed by the integration of registration-specific designs. The motion pyramid combined with warping (**DWP**) substantially enhances performance through coarse-to-fine refinement (+ $\sim$ 2% DSC), reaching the same level of performance as the Transformer baseline (TM) with greater efficiency. Adding a correlation volume (**DWCP**) further improves pixel-wise correspondence, yielding the best balance between alignment accuracy and deformation regularity (highest DSC with near-zero NDV). The explicit modeling of local similarity

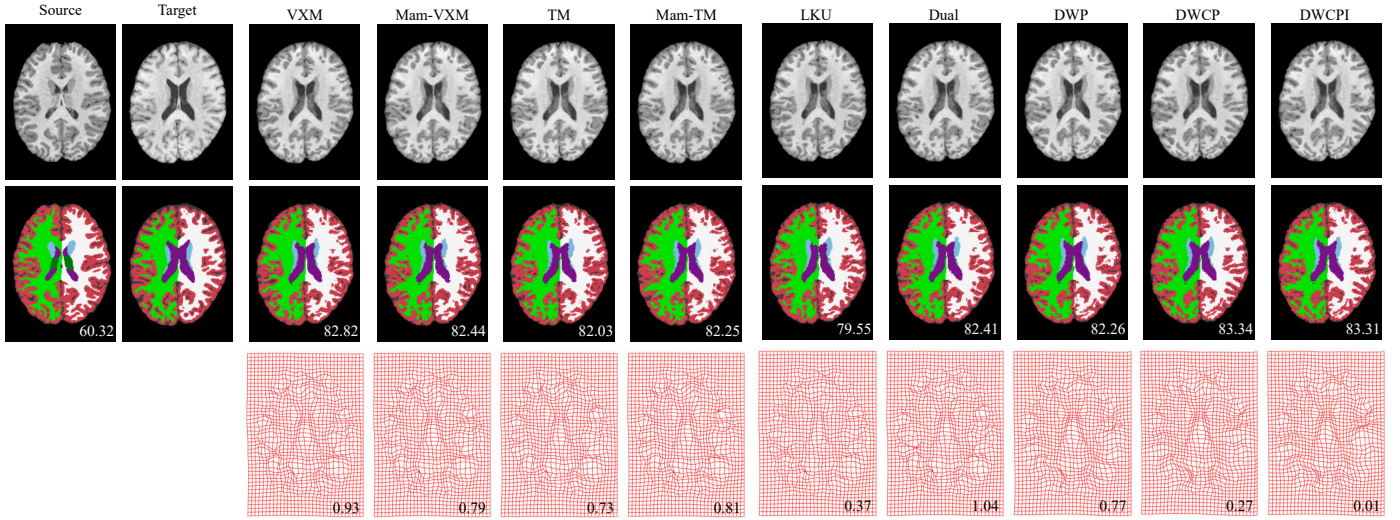


Fig. 3: Qualitative registration results on the LUMIR25 in-distribution (ID) dataset. The axial views of the source, target, and registered images are shown alongside their corresponding segmentation labels. The deformation field is illustrated as deformed grid lines. The Dice similarity coefficient (DSC, %) and the ratio of non-positive determinant voxels (NDV) across the volume are indicated in the bottom right corner.

Table 2: Quantitative results for cross-sectional brain registration on LUMIR 25 (In-Distribution (ID) (9 pairs) and Out-of-Distribution (OOD)) (18 pairs), LPBA (200 random pairs from 40 subjects), and MindBoggle (200 random pairs from 100 subjects). Accuracy is measured by Dice score (DSC, %), 95th percentile Hausdorff distance (HD95), Target Registration Error (TRE), Deformation regularity is quantified by standard deviation of log-Jacobian determinant (SDlogJ, $\times 10^{-2}$) and the ratio of non-diffeomorphic volume (NDV, %). The best (yellow, bold), second best (grey, italic), and third best (orange) results are highlighted.

	LUMIR25							LPBA				MindBoggle			
	ID			OOD											
	DSC $\uparrow$	HD95 $\downarrow$		DSC $\uparrow$	HD95 $\downarrow$	TRE $\downarrow$	NDV(%) $\downarrow$	DSC $\uparrow$	HD95 $\downarrow$	SDlogJ $\downarrow$	NDV(%) $\downarrow$	DSC $\uparrow$	HD95 $\downarrow$	SDlogJ $\downarrow$	NDV(%) $\downarrow$
Initial	56.68 $\pm$ 1.67	4.79 $\pm$ 0.24		51.89 $\pm$ 3.62	5.27 $\pm$ 0.61	4.35 $\pm$ 0.55	0.00 $\pm$ 0.00	54.0 $\pm$ 14.2	8.31 $\pm$ 3.46	-	-	51.5 $\pm$ 24.5	5.16 $\pm$ 3.88	-	-
VXM	74.35 $\pm$ 1.29	3.81 $\pm$ 0.21		70.26 $\pm$ 3.35	4.23 $\pm$ 0.62	2.64 $\pm$ 0.39	0.31 $\pm$ 0.10	67.1 $\pm$ 13.6	7.60 $\pm$ 3.48	7.2 $\pm$ 0.5	1.5 $\pm$ 0.7	67.2 $\pm$ 24.1	5.47 $\pm$ 4.75	7.6 $\pm$ 0.4	0.8 $\pm$ 0.2
Mam-VXM	74.02 $\pm$ 1.75	3.77 $\pm$ 0.21		70.27 $\pm$ 3.22	4.21 $\pm$ 0.59	2.65 $\pm$ 0.37	0.25 $\pm$ 0.08	68.1 $\pm$ 12.9	7.44 $\pm$ 3.38	6.7 $\pm$ 0.6	0.9 $\pm$ 0.4	67.5 $\pm$ 24.0	5.46 $\pm$ 4.75	7.4 $\pm$ 0.3	0.8 $\pm$ 0.1
TM	74.06 $\pm$ 1.55	3.74 $\pm$ 0.20		69.59 $\pm$ 2.80	4.14 $\pm$ 0.52	2.71 $\pm$ 0.37	0.23 $\pm$ 0.09	68.2 $\pm$ 12.8	7.42 $\pm$ 3.39	6.6 $\pm$ 0.5	0.9 $\pm$ 0.4	67.1 $\pm$ 24.1	5.46 $\pm$ 4.77	7.5 $\pm$ 0.4	0.8 $\pm$ 0.1
Mam-TM	74.01 $\pm$ 1.71	3.73 $\pm$ 0.20		69.26 $\pm$ 3.01	4.22 $\pm$ 0.55	2.68 $\pm$ 0.40	0.26 $\pm$ 0.10	67.8 $\pm$ 13.0	7.46 $\pm$ 3.42	7.0 $\pm$ 0.6	1.1 $\pm$ 0.5	66.9 $\pm$ 24.0	5.46 $\pm$ 4.75	7.7 $\pm$ 0.4	0.9 $\pm$ 0.2
LKU	74.35 $\pm$ 1.29	3.81 $\pm$ 0.21		70.26 $\pm$ 3.35	4.23 $\pm$ 0.62	2.64 $\pm$ 0.39	0.31 $\pm$ 0.10	65.8 $\pm$ 13.2	7.65 $\pm$ 3.37	6.2 $\pm$ 0.8	1.1 $\pm$ 1.0	64.8 $\pm$ 24.5	5.64 $\pm$ 4.78	7.5 $\pm$ 0.5	0.9 $\pm$ 0.2
Dual	74.37 $\pm$ 1.30	3.71 $\pm$ 0.20		70.14 $\pm$ 3.08	4.19 $\pm$ 0.57	2.73 $\pm$ 0.40	0.33 $\pm$ 0.08	67.3 $\pm$ 13.0	7.51 $\pm$ 3.38	7.1 $\pm$ 0.7	1.2 $\pm$ 0.5	67.1 $\pm$ 24.0	5.47 $\pm$ 4.74	7.8 $\pm$ 0.5	0.9 $\pm$ 0.1
DWP	76.44 $\pm$ 1.40	3.32 $\pm$ 0.19		<i>73.52 $\pm$ 1.90</i>	3.54 $\pm$ 0.37	2.51 $\pm$ 0.38	0.28 $\pm$ 0.07	70.8 $\pm$ 10.7	6.89 $\pm$ 3.05	6.3 $\pm$ 0.6	0.6 $\pm$ 0.1	<i>67.7 $\pm$ 24.1</i>	5.42 $\pm$ 4.78	7.5 $\pm$ 0.4	0.8 $\pm$ 0.1
DWCP	76.90 $\pm$ 1.51	<i>3.32 $\pm$ 0.20</i>		<i>73.39 $\pm$ 2.60</i>	<i>3.65 $\pm$ 0.54</i>	<i>2.42 $\pm$ 0.32</i>	<i>0.09 $\pm$ 0.04</i>	71.6 $\pm$ 10.0	6.78 $\pm$ 2.95	5.3 $\pm$ 1.2	<i>0.1 $\pm$ 0.1</i>	67.6 $\pm$ 24.0	5.40 $\pm$ 4.74	6.9 $\pm$ 0.4	<i>0.2 $\pm$ 0.1</i>
DWCPI	<i>76.88 $\pm$ 1.44</i>	<i>3.34 $\pm$ 0.17</i>		73.79 $\pm$ 2.57	3.65 $\pm$ 0.58	2.36 $\pm$ 0.32	0.01 $\pm$ 0.01	<i>71.3 $\pm$ 10.3</i>	6.85 $\pm$ 2.99	<i>5.6 $\pm$ 0.6</i>	0.0 $\pm$ 0.0	67.7 $\pm$ 23.9	<i>5.40 $\pm$ 4.77</i>	<i>7.2 $\pm$ 0.4</i>	0.0 $\pm$ 0.0

Table 3: Zero-shot cross-sectional brain MRI registration on OASIS (84 subjects), ADNI (43 subjects), and IXI (115 subjects) datasets, using 200 randomly sampled image pairs each. Metrics as in Table 2. The best (yellow, bold), second best (grey, italic), and third best (orange) results are highlighted.

	OASIS				ADNI				IXI			
	DSC $\uparrow$	HD95 $\downarrow$	SDlogJ $\downarrow$	NDV(%) $\downarrow$	DSC $\uparrow$	HD95 $\downarrow$	SDlogJ $\downarrow$	NDV(%) $\downarrow$	DSC $\uparrow$	HD95 $\downarrow$	SDlogJ $\downarrow$	NDV(%) $\downarrow$
initial	57.3 $\pm$ 22.4	4.20 $\pm$ 2.41	-	-	52.8 $\pm$ 22.1	4.79 $\pm$ 2.82	-	-	54.5 $\pm$ 23.7	4.87 $\pm$ 3.83	-	-
VXM	79.7 $\pm$ 17.7	5.14 $\pm$ 5.63	7.9 $\pm$ 0.5	0.9 $\pm$ 0.2	74.0 $\pm$ 18.7	4.64 $\pm$ 3.93	8.0 $\pm$ 0.4	1.1 $\pm$ 0.3	71.7 $\pm$ 21.3	5.21 $\pm$ 5.38	8.0 $\pm$ 0.4	1.1 $\pm$ 0.3
Mam-VXM	78.4 $\pm$ 17.9	5.14 $\pm$ 5.47	7.6 $\pm$ 0.4	0.8 $\pm$ 0.2	73.8 $\pm$ 18.7	4.66 $\pm$ 4.02	7.7 $\pm$ 0.3	0.9 $\pm$ 0.2	70.0 $\pm$ 21.6	5.31 $\pm$ 5.38	7.7 $\pm$ 0.4	0.8 $\pm$ 0.3
TM	78.1 $\pm$ 18.2	5.18 $\pm$ 5.65	<i>7.3 $\pm$ 0.4</i>	0.7 $\pm$ 0.2	73.2 $\pm$ 18.8	4.69 $\pm$ 4.05	7.6 $\pm$ 0.4	0.9 $\pm$ 0.3	70.2 $\pm$ 21.6	5.31 $\pm$ 5.48	<i>7.4 $\pm$ 0.4</i>	0.8 $\pm$ 0.3
Mam-TM	78.2 $\pm$ 18.1	5.15 $\pm$ 5.48	<i>7.3 $\pm$ 0.4</i>	0.8 $\pm$ 0.2	72.9 $\pm$ 18.9	4.70 $\pm$ 4.00	7.8 $\pm$ 0.5	1.1 $\pm$ 0.3	69.9 $\pm$ 21.8	5.30 $\pm$ 5.41	<i>7.3 $\pm$ 0.4</i>	0.8 $\pm$ 0.4
LKU	74.1 $\pm$ 20.1	5.29 $\pm$ 5.49	6.3 $\pm$ 0.4	<i>0.2 $\pm$ 0.1</i>	70.0 $\pm$ 20.0	4.84 $\pm$ 3.94	6.7 $\pm$ 0.4	<i>0.5 $\pm$ 0.3</i>	67.4 $\pm$ 22.6	5.43 $\pm$ 5.36	6.3 $\pm$ 0.5	<i>0.4 $\pm$ 0.3</i>
Dual	78.6 $\pm$ 18.1	5.19 $\pm$ 5.62	8.2 $\pm$ 0.5	1.1 $\pm$ 0.2	73.5 $\pm$ 19.0	4.66 $\pm$ 3.95	8.3 $\pm$ 0.4	1.2 $\pm$ 0.3	70.6 $\pm$ 21.7	5.29 $\pm$ 5.45	8.2 $\pm$ 0.4	1.1 $\pm$ 0.3
DWP	<i>80.0 $\pm$ 17.6</i>	<i>5.08 $\pm$ 5.53</i>	7.7 $\pm$ 0.5	0.9 $\pm$ 0.2	<i>75.2 $\pm$ 18.3</i>	<i>4.58 $\pm$ 4.03</i>	7.7 $\pm$ 0.4	0.9 $\pm$ 0.2	<i>72.4 $\pm$ 21.1</i>	5.19 $\pm$ 5.51	8.0 $\pm$ 0.6	1.0 $\pm$ 0.2
DWCP	<i>80.9 $\pm$ 17.0</i>	<i>5.03 $\pm$ 5.57</i>	9.6 $\pm$ 2.6	<i>0.3 $\pm$ 0.1</i>	<i>75.5 $\pm$ 18.1</i>	<i>4.55 $\pm$ 4.03</i>	<i>7.2 $\pm$ 0.4</i>	<i>0.4 $\pm$ 0.1</i>	<i>73.1 $\pm$ 20.9</i>	5.10 $\pm$ 5.40	7.6 $\pm$ 0.8	0.5 $\pm$ 0.2
DWCPI	81.1 $\pm$ 16.8	5.03 $\pm$ 5.56	10.2 $\pm$ 3.1	0.1 $\pm$ 0.0	75.6 $\pm$ 18.0	4.54 $\pm$ 4.01	7.4 $\pm$ 0.4	0.0 $\pm$ 0.0	73.2 $\pm$ 20.7	<i>5.10 $\pm$ 5.42</i>	7.7 $\pm$ 0.5	0.0 $\pm$ 0.0

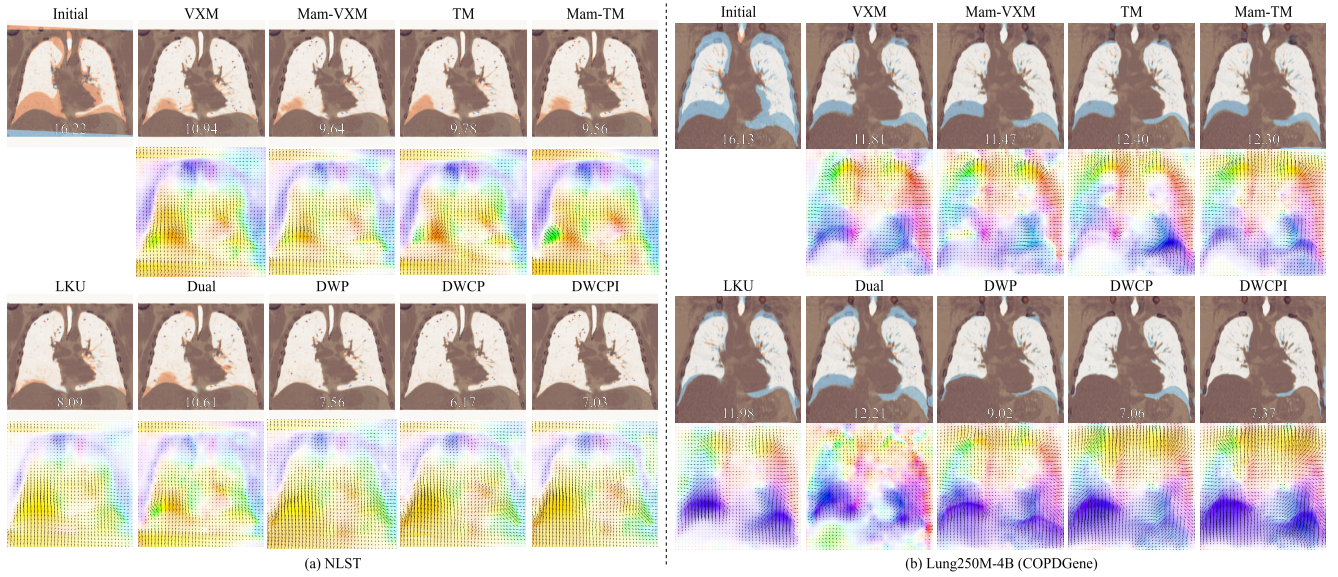


Fig. 4: Qualitative registration results on the NLST (in-domain) and Lung250M-4B (zero-shot) datasets. The coronal overlays of the registered moving and target images are shown alongside their corresponding deformation fields. The target and source images are visualized in **blue** and **organ**, and a perfectly aligned lung appears **white** in the overlay. The deformation field is visualized by a colored quiver plot where the color and arrow direction encode the displacement orientation, and arrow length represents displacement magnitude. The target registration error (TRE) for each case is indicated at the bottom of the image.

Table 4: Quantitative results on the NLST (in-domain) and Lung250M-4B (zero-shot) datasets. Reported metrics include target registration error (TRE, mm or vox), Dice score (DSC, %), and deformation regularity measures: SDlogJ, ($\times 10^{-2}$); and the percentage of non-diffeomorphic voxels (NDV) across the image volume. The best (**, bold**), second best (*, italic*), and third best (**, bold**) results are highlighted.

	NLST (In-domain)								Lung250M-4B (Zero-shot)							
	INP → EXP				EXP → INP				INP → EXP				EXP → INP			
	TRE ↓	DSC ↑	SDlogJ ↓	NDV(%) ↓	TRE ↓	DSC ↑	SDlogJ ↓	NDV(%) ↓	TRE ↓	SDlogJ ↓	NDV(%) ↓	TRE ↓	SDlogJ ↓	NDV(%) ↓		
Initial	10.14±2.92	89.2±3.0	-	-	10.14±2.92	89.2±3.0	-	-	13.74±3.63	-	-	13.74±3.63	-	-		
VXM	3.66±2.60	97.5±2.1	7.4±1.3	0.58±0.78	3.62±2.59	97.6±1.9	7.1±1.0	0.48±0.55	8.98±4.52	10.5±2.8	0.72±0.65	8.86±4.38	9.1±1.9	0.66±0.69		
Mam-VXM	3.47±2.60	97.6±2.1	7.7±1.4	0.66±0.91	3.42±2.47	97.7±1.9	7.3±1.0	0.43±0.40	9.18±4.77	11.3±3.4	1.05±0.99	8.89±4.53	9.5±2.1	0.68±0.96		
TM	3.41±2.53	97.6±2.0	7.6±1.2	0.55±0.63	3.46±2.61	97.6±2.1	7.4±1.1	0.65±0.80	9.38±4.94	11.4±3.5	1.11±0.97	8.87±4.60	9.3±2.1	0.66±0.60		
Mam-TM	3.42±2.46	97.6±1.9	7.4±1.3	0.48±0.56	3.38±2.46	97.7±1.8	7.1±1.0	0.38±0.38	9.06±4.72	11.0±3.6	1.08±1.01	8.88±4.51	9.2±2.2	0.82±0.86		
LKU	3.78±1.87	97.6±1.3	6.7±1.4	0.37±0.58	3.75±1.78	97.6±1.1	6.6±1.2	0.19±0.24	8.69±4.27	8.6±2.6	0.21±0.27	7.64±3.75	7.7±1.8	0.18±0.22		
Dual	4.41±2.37	97.4±2.1	7.8±1.3	0.50±0.53	4.42±2.48	97.5±2.1	7.5±1.0	0.44±0.48	11.27±4.12	11.6±1.9	0.99±0.75	11.23±4.13	10.4±1.6	0.83±0.70		
DWP	2.60±1.44	98.4±0.7	7.9±1.5	0.31±0.38	2.60±1.41	98.4±0.6	7.9±1.5	0.20±0.16	9.64±4.22	13.1±2.7	0.52±0.35	9.51±4.16	10.9±2.1	0.52±0.39		
DWCP	2.00±1.45	98.4±1.0	7.5±1.2	0.13±0.12	2.03±1.46	98.4±0.7	7.6±1.5	0.11±0.10	8.04±4.02	12.7±3.1	0.17±0.12	7.58±3.91	9.5±1.7	0.13±0.11		
DWCPI	1.95±1.47	98.4±0.9	7.6±1.1	0.05±0.06	2.00±1.49	98.4±0.6	7.7±1.3	0.04±0.05	8.42±4.37	12.6±2.7	0.06±0.07	7.88±4.31	9.6±1.4	0.02±0.02		

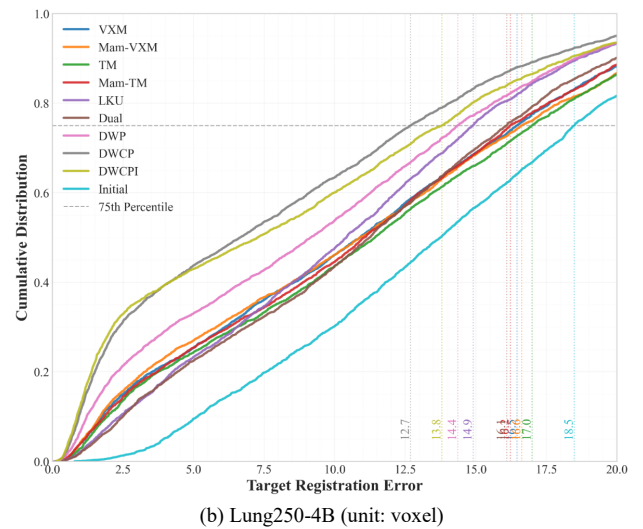
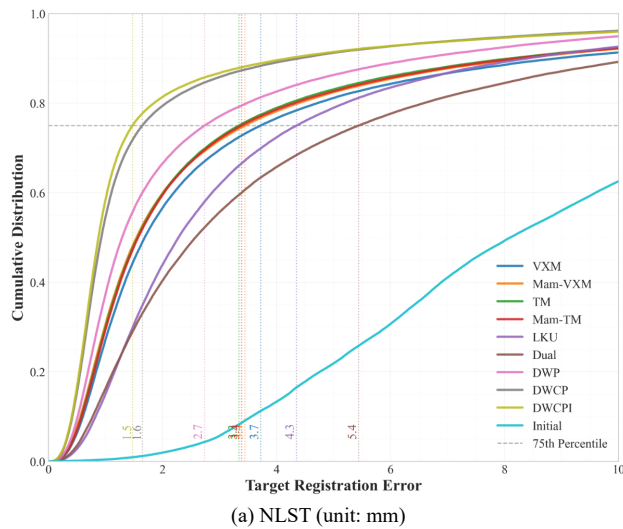


Fig. 5: Cumulative distribution of target registration error (TRE) on the NLST (in-domain) and Lung250M-4B (zero-shot) datasets. The dotted vertical lines indicate the 75th percentile of TRE in each distribution. A larger area under the curve corresponds to better keypoints/landmarks alignment performance.

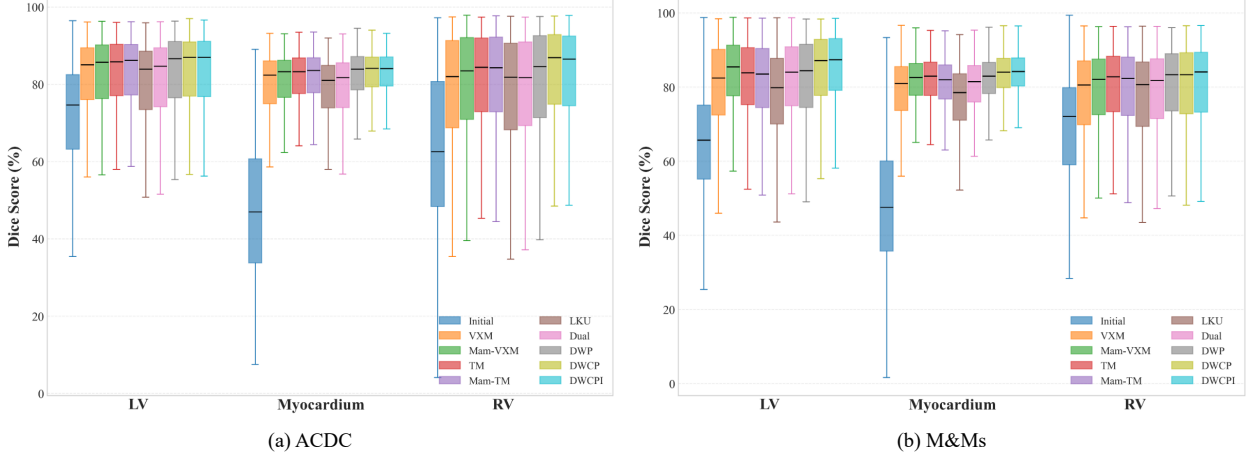


Fig. 6: Box plots of structure-wise Dice scores on the ACDC (in-domain) and M&Ms (zero-shot) datasets. Structures include the left ventricle (LV), myocardium (Myo), and right ventricle (RV).

Table 5: Quantitative results on the ACDC (in-domain) and M&Ms (zero-shot) test sets. Reported metrics include Dice score (DSC,%), Normalized Surface Dice (NSD,%), the standard deviation of the log-Jacobian determinant ($\text{SDlogJ} \times 10^{-2}$), and the number of non-diffeomorphic voxels within the foreground (NDV,%_{cc}, per 10,000 voxels). The best (●, **bold**), second best (●, *italic*), and third best (●) results are highlighted.

	ACDC								M&Ms (Zero-shot)							
	ED → ES				ES → ED				ED → ES				ES → ED			
	DSC ↑	NSD ↓	SDlogJ ↓	NDV(% _{cc}) ↓	DSC ↑	NSD ↓	SDlogJ ↓	NDV(% _{cc}) ↓	DSC ↑	NSD ↓	SDlogJ ↓	NDV(% _{cc}) ↓	DSC ↑	NSD ↓	SDlogJ ↓	NDV(% _{cc}) ↓
initial	60.2±14.8	10.7±7.3	-	-	60.2±14.8	10.7±7.3	-	-	60.4±12.7	7.9±6.0	-	-	60.4±12.7	7.9±6.0	-	-
VXM	78.8±10.8	22.4±6.2	4.7±1.6	0.3±1.8	83.9±8.3	20.8±7.2	4.1±1.2	3.0±13.3	77.7±11.0	15.1±6.3	4.2±1.8	16.8±40.5	83.8±7.8	18.8±7.7	3.4±1.0	15.5±34.9
Mam-VXM	80.0±9.9	22.8±6.1	5.1±1.9	1.5±5.6	85.3±7.6	22.5±7.4	4.4±1.5	7.0±21.0	80.4±8.5	16.3±6.2	4.1±1.7	29.9±60.8	85.0±7.1	20.2±7.6	3.5±1.4	48.9±88.4
TM	80.4±9.7	23.2±5.9	5.0±1.9	1.1±5.1	85.5±7.4	22.3±7.2	4.3±1.4	4.4±17.9	80.2±9.1	17.9±6.4	3.4±1.5	1.9±6.8	84.8±7.5	17.9±8.2	2.9±1.2	3.2±9.1
Mam-TM	80.4±9.9	22.9±6.2	5.1±1.9	1.2±5.4	85.4±7.7	22.3±7.1	4.3±1.4	5.4±18.6	79.5±9.3	18.0±6.6	3.4±1.4	9.2±26.8	84.6±7.4	17.5±7.9	2.9±1.1	10.3±25.0
LKU	77.7±10.7	21.6±6.4	4.5±1.9	0.1±0.7	82.2±8.7	18.0±6.7	3.8±1.4	5.3±19.7	76.4±10.3	16.3±6.2	3.0±1.4	0.3±2.0	80.5±9.5	14.5±7.3	2.5±1.1	2.4±7.2
Dual	78.1±11.0	22.3±5.6	4.7±1.5	0.2±1.7	83.1±8.8	19.7±6.7	3.9±1.1	2.7±12.8	79.3±9.5	17.4±6.1	3.4±1.1	3.1±14.2	83.6±7.7	17.4±7.7	2.9±1.1	3.5±11.9
DWP	80.3±10.9	23.3±5.7	4.9±1.6	0.1±0.6	85.4±8.5	23.2±7.5	4.3±1.3	0.7±3.9	80.3±9.5	18.8±6.3	3.5±1.2	0.7±3.6	85.3±7.6	18.8±7.8	3.2±1.2	0.8±3.3
DWCP	81.3±10.3	23.9±5.7	5.2±3.5	0.5±4.1	86.1±7.9	24.4±7.8	4.4±1.3	1.0±5.1	81.3±9.6	19.0±6.3	3.9±3.5	1.6±9.5	86.2±7.5	20.3±7.5	3.5±3.7	1.6±8.6
DWCPi	<i>81.2±10.4</i>	23.8±5.7	5.0±1.7	0.3±2.5	<i>85.8±8.4</i>	<i>23.6±7.9</i>	4.4±1.3	0.2±1.9	81.7±9.5	19.0±6.3	4.3±5.3	1.3±7.8	86.2±7.4	20.1±7.5	3.9±5.3	<i>1.2±7.2</i>

via the correlation enables more precise delineation of ventricular boundaries and myocardial contours (Figure 7), particularly in regions with limited texture contrast. This advantage underscores that local correlation remains a powerful inductive bias, surpassing computation blocks capable of global context modeling.

While the iterative refinement variant (**DWCPi**) gets modest gains over **DWCP** on accuracy, qualitative inspection in Figure 7 reveals that **DWCPi** achieves more accurate alignment of RV in the difficult M&Ms case with significant shape changes. This highlights the potential of iterative refinement to enhance robustness in extreme motion scenarios.

Last but not least, as illustrated by the box plots in Figure 6, registration-specific designs consistently yield higher and more stable DSC across all cardiac structures (LV, Myo, RV), confirming that the explicit registration priors play a decisive role in achieving cross-domain stability and anatomical consistency.

5.4. Intra-subject Abdomen MRI-CT Registration

Table 6 and Figure 10 present both quantitative and qualitative results for intra-subject MRI-CT abdominal registration, one of the most challenging settings in our benchmark. Despite being intra-subject, the task involves substantial variations caused by changes in patient posture, field-of-view differences,

and potential organ resection. Moreover, the contrast difference of the MRIs casts a significant domain shift for the multi-modal registration. These factors contribute to making this task a robust test of model generalization and robustness.

Consistent with other tasks, the low-level trend-driven architectures offer limited improvement over the VXM baseline. As shown in Figure 9, Transformer- and Mamba-based variants fail to significantly improve alignment accuracy across organs, suggesting that their general capacity for long-range modeling might not translate into effective cross-modal spatial matching. The LKU-Net model even underperforms the initial alignment on the spleen and right kidney, which could be attributed to the fact that the enlarged receptive field might not be helpful for the significant organ shifting.

In contrast, registration-specific designs, particularly the correlation volume combined with the motion pyramid (**DWCP**), yield the most substantial and consistent performance gains across all metrics (e.g., +4% DSC, +0.9% NSD). This demonstrates that explicitly computing modality-agnostic similarity via latent feature correlation is one of the keys to bridging the intensity gap between MRI and CT. Interestingly, the motion pyramid without the correlation volume (**DWP**) only achieves on-par performance with the VXM baseline and significantly degrades deformation smoothness.

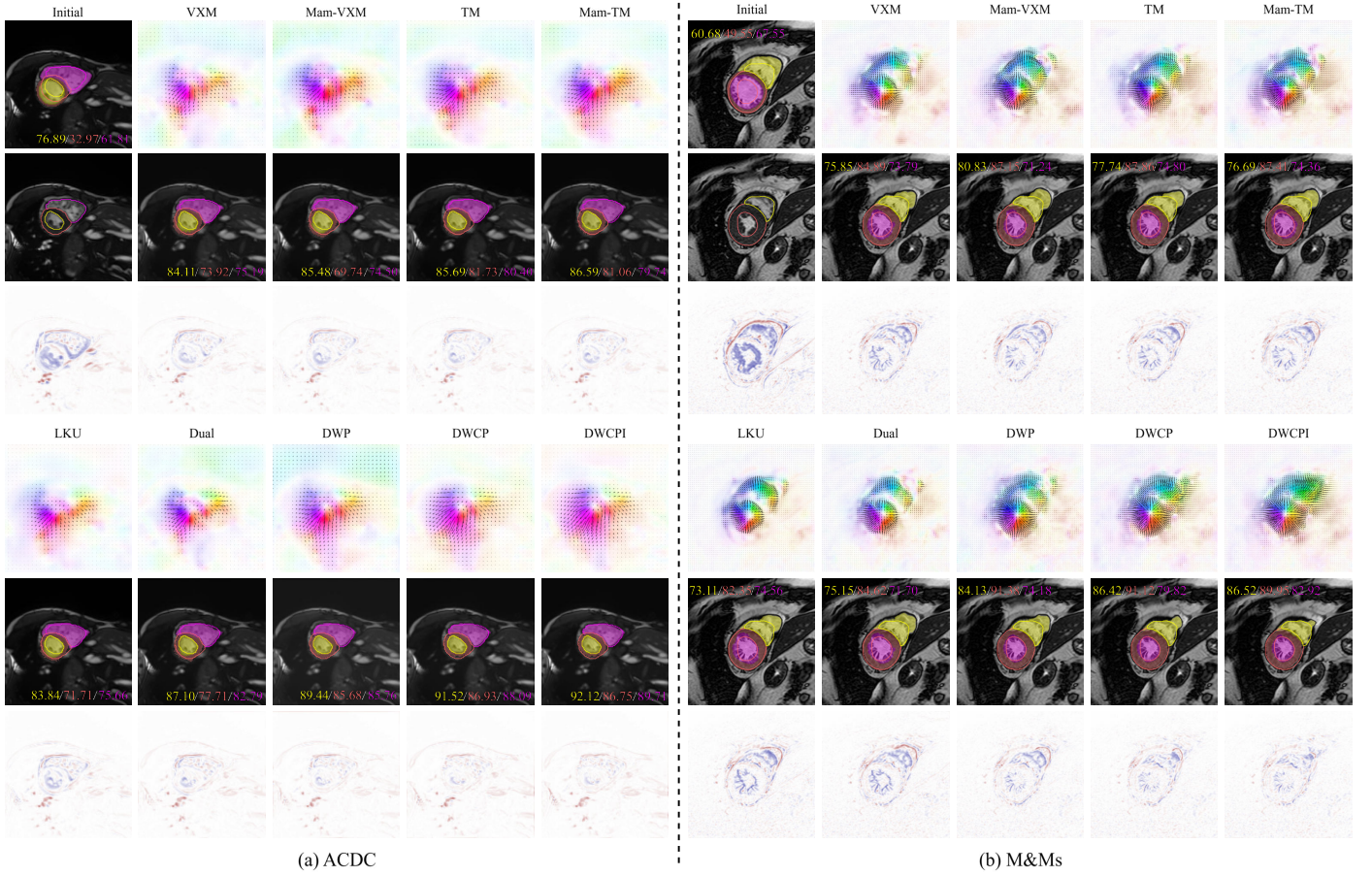


Fig. 7: Qualitative registration results on the ACDC and M&Ms dataset. The target image is shown with its segmentation contour overlaid. The source and registered images are overlaid with both their respective segmentation labels and the target contour. Subtraction error maps between the target and source/registered images are displayed, where white denotes zero intensity difference. Perfect registration aligns the contour precisely with the segmentation boundary. Dice scores for the left ventricle, myocardium, and right ventricle are reported within each image. The deformation field is visualized by a colored quiver plot, where the color and arrow direction encode the displacement orientation, and arrow length represents displacement magnitude.

Table 6: Quantitative results for the Abdomen MRI-CT registration task. Reported metrics include Dice score (DSC,%), Normalized Surface Dice (NSD,%), the standard deviation of the log-Jacobian determinant (SDlogJ, $\times 10^{-2}$), and the number of non-diffeomorphic voxels within the foreground (NDV, $\%_{cc}$, per 10,000 voxels). The best (■, **bold**), second best (■, *italic*), and third best (■) results are highlighted.

	MRI → CT				CT → MRI			
	DSC ↑	NSD ↑	SDlogJ ↓	NDV($\%_{cc}$) ↓	DSC ↑	NSD ↑	SDlogJ ↓	NDV($\%_{cc}$) ↓
affine	64.9±8.2	8.6±2.3	-	-	64.9±8.2	8.6±2.3	-	-
VXM	69.0±3.9	11.1±1.8	9.2±1.3	0.2±0.3	68.9±4.5	12.5±1.2	9.6±1.6	3.6±4.1
Mam-VXM	68.5±3.8	11.1±1.7	9.8±0.8	3.6±9.5	67.6±4.2	12.4±1.3	10.0±1.4	29.1±32.9
TM	68.4±3.2	11.0±2.1	10.1±0.9	3.4±4.7	67.9±3.9	12.9±1.3	9.8±1.3	30.0±30.7
Mam-TM	69.6±3.5	11.0±1.9	10.2±1.0	21.9±56.6	67.5±4.1	12.5±1.5	9.8±1.2	41.8±52.1
LKU	65.0±6.1	10.9±1.8	7.5±2.1	0.0±0.0	63.8±8.4	11.3±1.6	7.1±2.3	0.1±0.2
Dual	67.6±5.6	11.2±1.2	10.1±1.0	2.9±3.5	67.3±5.4	12.5±1.2	10.0±1.3	15.0±21.3
DWP	69.0±3.7	10.9±1.9	11.1±2.2	9.9±16.2	69.6±4.6	13.3±1.6	11.2±1.5	57.8±41.9
DWCP	73.4±6.8	11.8±2.1	10.3±1.6	1.0±2.0	72.9±6.9	13.4±1.3	10.0±1.8	7.6±7.0
DWCPi	72.2±7.2	11.1±2.1	9.9±1.6	0.0±0.0	71.5±7.7	13.3±1.5	9.3±1.3	0.0±0.0

Finally, the results for the iterative refinement (**DWCPi**) reveal a critical trade-off. While it successfully regularizes the deformation field, reaching lower SDlogJ and near-zero folding voxels, this comes at the cost of a slight drop in registration accuracy compared to the single-shot **DWCP**. As illustrated in the top-right subject of Figure 10, the iterative refinement leads to poorer alignment of the kidneys, suggesting that the recurrent

update may converge toward a suboptimal deformation.

5.5. Progressive Deformation Refinement Analysis

To better understand the cumulative benefits of our registration-specific components, namely, the motion pyramid, correlation volume, and iterative refinement, we perform a fine-grained analysis of registration accuracy across all refinement

Progressive Refinement via Coarse-to-Fine Pyramid and Iteration

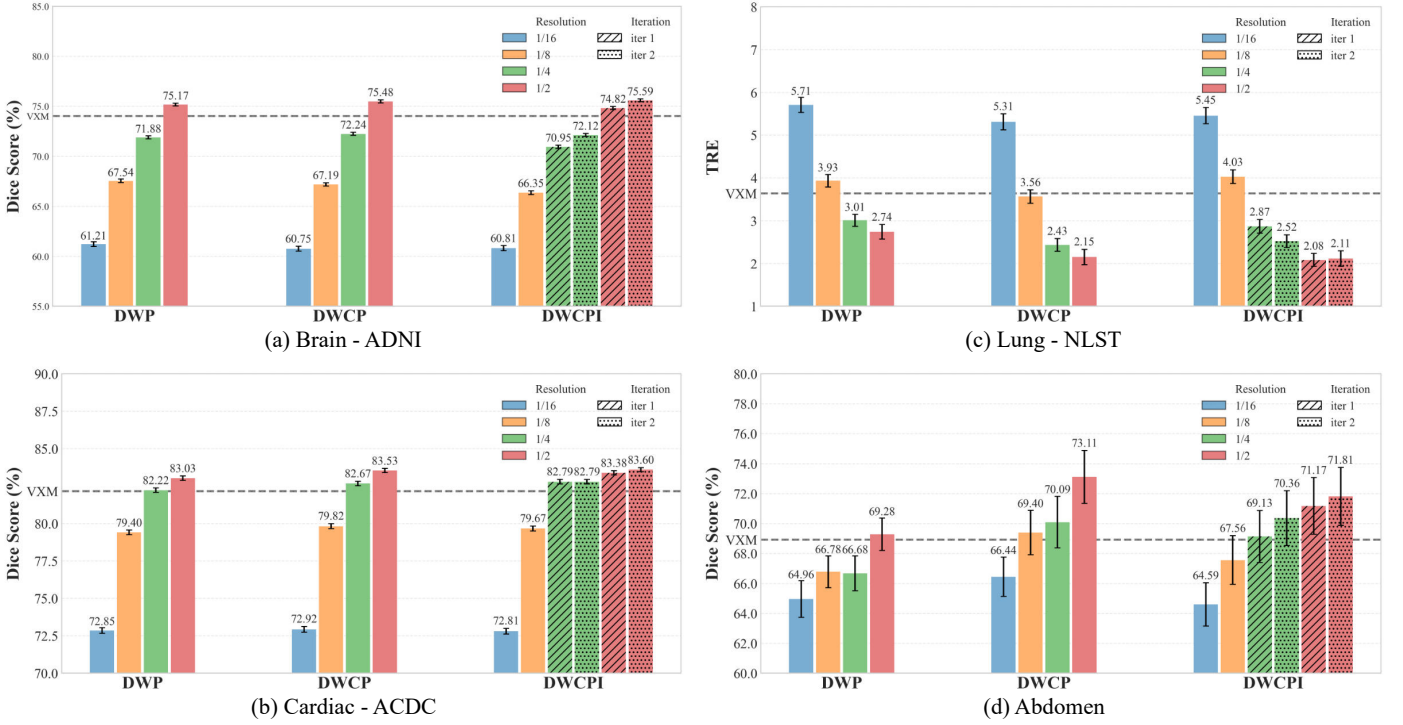


Fig. 8: Progressive deformation refinement analysis across all tasks. Each subplot shows the incremental improvements in registration accuracy through coarse-to-fine pyramid refinement and iterative updates: (a) brain (ADNI), (b) cardiac (ACDC), (c) lung (NLST), and (d) abdomen anatomies. The VXIM baseline reference is marked by the dotted horizontal line. Tables of additional metrics are provided in Section [Appendix A.1](#).

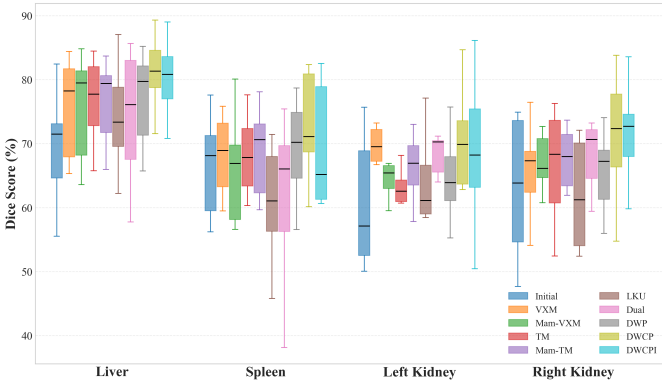


Fig. 9: Box plots of structure-wise Dice scores on the Abdomen CT-MRI registration task. Structures include liver, spleen, left kidney and right kidney. A detailed breakdown of the Dice scores for each subject and organ is provided in Section [Appendix A.2](#).

steps. For each variant, we evaluate metrics at every resolution level and iterative step to disentangle how these modules individually and collectively contribute to the final performance.

Motion pyramid achieves near-optimal alignment at surprisingly coarse scales. As shown in Figure 8, for most tasks and model variants, the registration accuracy at 1/4 resolution already matches or even surpasses the VXIM baseline (indicated by the VXIM reference line crossing green bars). This demonstrates that a pyramidal design effectively captures large-scale deformations early, requiring only two or three coarse-to-fine

refinements to achieve an anatomically meaningful alignment. In essence, most of the “heavy lifting” in deformation estimation happens at surprisingly low resolutions.

Correlation layer provides a strong inductive bias for voxel correspondence. When comparing **DWP** and **DWCP**, a clear and consistent trend emerges: adding the correlation layer boosts accuracy across all pyramid levels, with the largest gains observed at resolutions 1/4 and 1/2. This improvement highlights the importance of explicitly encoding local spatial similarity: correlation volumes act as an interpretable prior that reinforces voxel-to-voxel matching, complementing the implicit correspondence learned by the deformation decoder.

Iterative refinement improves fine details—but encourages “lazy” decoders. An interesting phenomenon arises when introducing iterative refinement (**DWCPi**). While iterations indeed improve the final accuracy at finer resolutions (1/4, 1/2), as evidenced by the dashed versus dotted bars of the same color, they often lead to slightly reduced performance at coarser scales (1/16, 1/8) compared to the single-shot **DWCP**. This suggests a “lazy decoder” behavior: once the network learns that an iterative update will follow, the early-stage deformation estimations relax their effort, deferring the correction to later iterations. This effect can be partially attributed to the gradient flow through the unrolled iteration, similar to recurrent networks, where accumulated gradients bias learning at a coarser level. This insight opens an interesting avenue for future research: en-

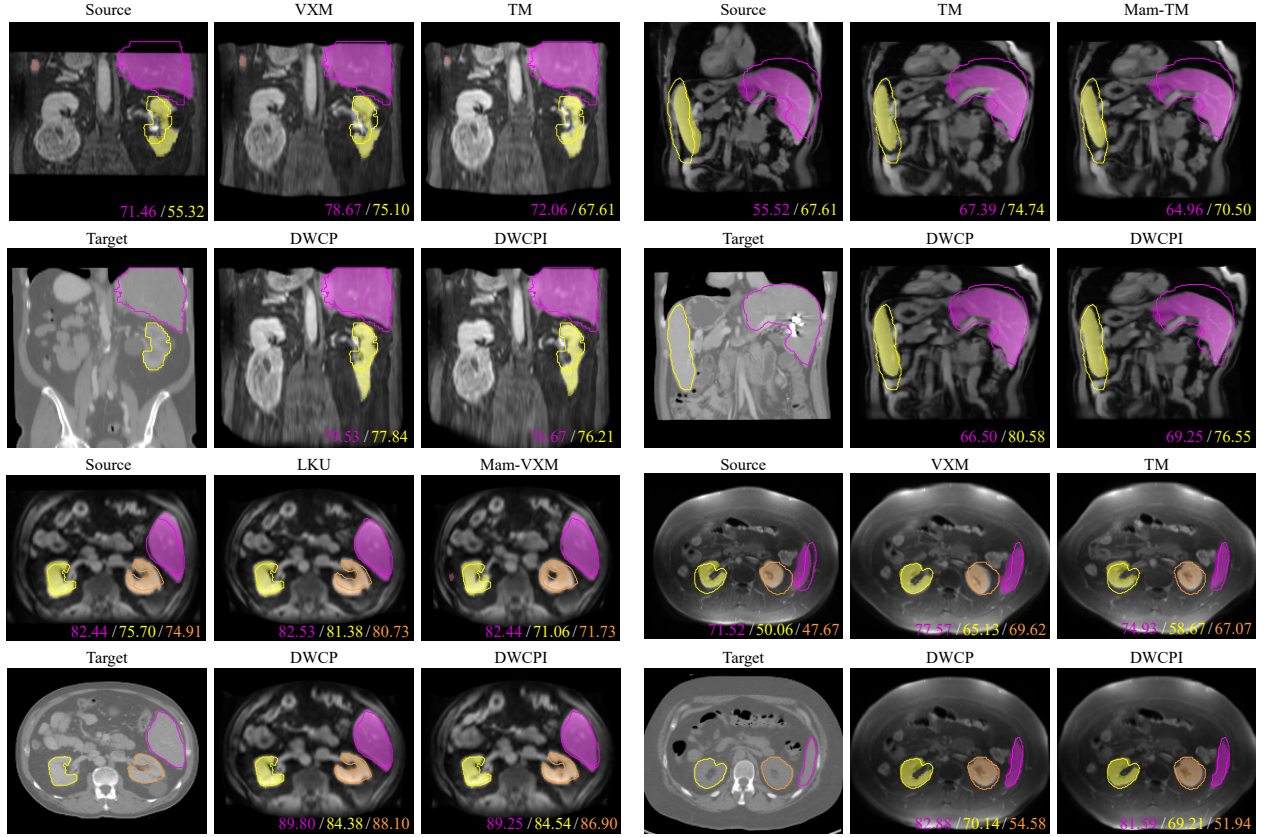


Fig. 10: Qualitative registration results on Abdomen MRI-CT registration. The target image is shown with its segmentation contour overlaid. The source and registered images are overlaid with both their respective segmentation labels and the target contour. Perfect registration aligns the contour precisely with the segmentation boundary. Dice scores for the liver, spleen, left/right kidney are reported within each image.

couraging each resolution to reach near-optimal alignment before iterative refinement begins.

Summary. Overall, this progressive refinement analysis provides concrete evidence that domain-specific architectural priors play a decisive role in improving deformable registration. The motion pyramid captures coarse global motion efficiently, the correlation layer enhances correspondence precision, and the iterative inference refines localized structures. Together, they form a principled hierarchy of complementary mechanisms, each contributing uniquely to robust and accurate deformation estimation across diverse imaging domains.

6. Conclusion and Discussion

In this work, we revisited a fundamental question in learning-based deformable image registration: what truly drives progress? Through a modular framework and comprehensive experiments across four diverse tasks, covering brain, lung, cardiac, and abdominal registration under various imaging modalities, we systematically disentangled the influence of low-level computation blocks from high-level, registration-specific designs.

Our findings consistently demonstrate that progress in registration is rooted in domain-specific inductive priors, rather than adopting “more advanced” or “trend-driven” architectures.

While modern computational blocks such as Transformers, Mamba, or large-kernel convolutions offer marginal improvements, they come with substantial computational overhead and limited generalization. In contrast, registration-specific designs, including dual-stream encoder, motion pyramid, correlation, and iterative refinement, consistently yield more accurate, smoother, and more robust deformations across modalities and domains.

The correlation-based motion pyramid emerges as a particularly effective and generalizable design, bridging large displacement estimation and cross-modal matching. Meanwhile, the analysis of iterative refinement reveals both its promise and pitfalls, providing an intriguing research direction on balancing hierarchical and recursive optimization within a single model.

Beyond empirical findings, we release our modular, open-source benchmark as a foundation for fair, transparent, and extensible evaluation. This framework allows future studies to isolate the genuine contribution of new methods, ensuring that reported gains are not confounded by hidden priors or inconsistent training setups. We envision this as a living benchmark, continuously evolving with community contributions. (<https://github.com/BailiangJ/rethink-reg>)

Limitations. While our study focuses on representative architectures and registration-specific designs, many paradigms imported from computer vision, such as diffusion-based registration, implicit neural representations (INR)-based deformations,

or task-specific strategies like B-spline parameterization, remain unexplored. Nevertheless, these approaches can be seamlessly integrated into our modular benchmark. We view this work as an essential milestone toward a unified, extensible framework for fair and reproducible registration research. Besides, we do not carry out extensive hyperparameter, architectural, task-specific tuning to ensure fair and reproducible comparison across all methods. While this may not yield the absolute optimal performance for each method, it allows unbiased assessment of design principles and preserves the validity of our findings.

In summary, our results advocate a shift in research emphasis: from chasing architectural trends to understanding and leveraging domain priors. By grounding future work in principled, reproducible analysis, we hope to foster more transparent, generalizable, and clinically meaningful progress in medical image registration.

Acknowledgement. This work was supported by BMWi (project “NeuroTEMP”) research funding and the Munich Center of Machine Learning (MCML). We gratefully thank the computational resources provided by the Leibniz Supercomputing Centre (LRZ).

Disclosure of Interests. The authors declare no competing interests.

Declaration of usage of LLMs. During the preparation of this work, the authors used ChatGPT to improve its readability. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

- Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C., 2008. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical image analysis* 12, 26–41.
- Baheti, B., Waldmannstetter, D., Chakrabarty, S., Akbari, M., Bilello, M., Wiestler, B., Schwarting, J., Calabrese, E., Jeffrey, R., Abidi, S., et al., 2021. The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up mri scans of diffuse glioma patients. *ArXiv.org*.
- Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V., 2019. Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* 38, 1788–1800.
- Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: *European conference on computer vision*, Springer. pp. 404–417.
- Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., Cetin, I., Lekadir, K., Camara, O., Ballester, M.A.G., et al., 2018. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging* 37, 2514–2525.
- Brock, K.K., Mutic, S., McNutt, T.R., Li, H., Kessler, M.L., 2017. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132. *Medical physics* 44, e43–e76.
- Brox, T., Bruhn, A., Papenberg, N., Weickert, J., 2004. High accuracy optical flow estimation based on a theory for warping, in: *Computer Vision-ECCV 2004*, Springer. pp. 25–36.
- Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al., 2021. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging* 40, 3543–3554.
- Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y., 2022. Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis* 82, 102615.
- Chen, J., He, Y., Frey, E.C., Li, Y., Du, Y., 2021. Vit-v-net: Vision transformer for unsupervised volumetric medical image registration. *arXiv preprint arXiv:2104.06468*.
- Chen, J., Wei, S., Honkamäa, J., Marttinen, P., Zhang, H., Liu, M., Zhou, Y., Tan, Z., Wang, Z., Wang, Y., et al., 2025a. Beyond the lumir challenge: The pathway to foundational registration models. *arXiv preprint arXiv:2505.24160*.
- Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K., 2018. Neural ordinary differential equations. *Advances in neural information processing systems* 31.
- Chen, X., Liu, X., Wang, Z., Wei, Y., 2024. Multi-scale lung ct image registration based on mamba-convnet, in: *2024 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE. pp. 1–5.
- Chen, Y., Wang, X., Qian, C., Guan, X., Dai, X., Gao, P., 2025b. Mambabir: A residual pyramid network for brain image registration with state-space model. *Expert Systems with Applications*, 129335.
- Chen, Z., Zheng, Y., Gee, J.C., 2023. Transmatch: A transformer-based multi-tile dual-stream feature matching network for unsupervised deformable image registration. *IEEE transactions on medical imaging* 43, 15–27.
- Cheng, X., Zhang, T., Lu, W., Meng, Q., Frangi, A.F., Duan, J., 2025. Sacbnet: Spatial-awareness convolutions for medical image registration, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5227–5237.
- Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al., 2013. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging* 26, 1045–1057.
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks, in: *Proceedings of the IEEE international conference on computer vision*, pp. 764–773.
- Ding, X., Zhang, X., Han, J., Ding, G., 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11963–11975.
- Dorent, R., Rigolo, L., Galvin, C.P., Chen, J., Heinrich, M.P., Carass, A., Colliot, O., Wassermann, D., Golby, A., Kapur, T., et al., 2025. The brain resection multimodal image registration (remind2reg) 2025 challenge. *arXiv preprint arXiv:2508.09649*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
- Dosovitskiy, A., Fischer, P., Ilg, E., et al., 2015. FlowNet: Learning optical flow with convolutional networks, in: *Proceedings of the IEEE international conference on computer vision*, pp. 2758–2766.
- Falta, F., Großbröhm, C., Hering, A., Bigalke, A., Heinrich, M., 2023. Lung250m-4b: a combined 3d dataset for ct-and point cloud-based intrapatient lung registration. *Advances in Neural Information Processing Systems* 36, 54819–54832.
- Ghahremani, M., Khateri, M., Jian, B., Wiestler, B., Adeli, E., Wachinger, C., 2024. H-vit: A hierarchical vision transformer for deformable image registration, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11513–11523.
- Greer, H., Kwitt, R., Vialard, F.X., Niethammer, M., 2021. Icon: Learning regular maps through inverse consistency, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3396–3405.
- Greer, H., Tian, L., Vialard, F.X., Kwitt, R., Bouix, S., San Jose Estepar, R., Rushmore, R., Niethammer, M., 2023. Inverse consistency by construction for multistep deep registration, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 688–698.
- Gu, A., Dao, T., 2024. Mamba: Linear-time sequence modeling with selective state spaces. *International Conference on Machine Learning (ICML)*.
- Guo, T., Wang, Y., Meng, C., 2024. Mambamorph: a mamba-based backbone with contrastive feature learning for deformable mr-ct registration. *arXiv preprint arXiv:2401.13934*.
- Hansen, L., Heyer, W., Großbröhm, C., Madesta, F., Sentker, T., Jiazheng, W., Zhang, Y., Zhang, H., Liu, M., Wang, J., et al., 2025. Learn2reg 2024: New benchmark datasets driving progress on new challenges. *arXiv preprint arXiv:2509.01217*.
- van Harten, L., Van Herten, R.L.M., Stoker, J., Isgum, I., 2024. Deformable

- image registration with geometry-informed implicit neural representations, in: *Medical Imaging with Deep Learning*, PMLR. pp. 730–742.
- Heinrich, M.P., 2019. Closing the gap between deep and conventional image registration using probabilistic dense displacement networks, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 50–58.
- Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A., 2013. Towards realtime multimodal fusion for image-guided interventions using self-similarities, in: *International conference on medical image computing and computer-assisted intervention*, Springer. pp. 187–194.
- Hering, A., Hansen, L., Mok, T.C., Chung, A.C., et al., 2022. Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging* 42, 697–712.
- Honkamaa, J., Martinen, P., 2023. Sitreg: Multi-resolution architecture for symmetric, inverse consistent, and topology preserving image registration. *arXiv preprint arXiv:2303.10211*.
- Hosni, A., Rhemann, C., others, 2012. Fast cost-volume filtering for visual correspondence and beyond. *IEEE transactions on pattern analysis and machine intelligence* 35, 504–511.
- Hu, X., Chen, J., Chen, Y., 2024. Regmamba: An improved mamba for medical image registration. *Electronics* (2079-9292) 13.
- Huang, T., Mu, N., Xing, H., Li, X., 2024. 3d deformable convolution for medical image registration, in: *Pacific Rim International Conference on Artificial Intelligence*, Springer. pp. 179–191.
- Im, J.E., Khalifa, M., Gregory, A.V., Erickson, B.J., Kline, T.L., 2025. A systematic review on the use of registration-based change tracking methods in longitudinal radiological images. *Journal of Imaging Informatics in Medicine* 38, 2549–2562.
- Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* 18, 203–211.
- Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F., 2024. nnu-net revisited: A call for rigorous validation in 3d medical image segmentation, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 488–498.
- Jack Jr, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., et al., 2008. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 27, 685–691.
- Jena, R., Chaudhari, P., Gee, J.C., 2024a. Fireants: Adaptive riemannian optimization for multi-scale diffeomorphic registration. *arXiv preprint arXiv:2404.01249*.
- Jena, R., Sethi, D., Chaudhari, P., Gee, J., 2024b. Deep learning in medical image registration: Magic or mirage? *Advances in Neural Information Processing Systems* 37, 108331–108353.
- Jia, X., Bartlett, J., Zhang, T., Lu, W., Qiu, Z., Duan, J., 2022. U-net vs transformer: Is u-net outdated in medical image registration?, in: *International Workshop on Machine Learning in Medical Imaging*, Springer. pp. 151–160.
- Jia, X., Thorley, A., Chen, W., Qiu, H., Shen, L., Styles, I.B., Chang, H.J., Leonardi, A., De Marvao, A., O'Regan, D.P., et al., 2021. Learning a model-driven variational network for deformable image registration. *IEEE Transactions on Medical Imaging* 41, 199–212.
- Jian, B., Pan, J., Ghahremani, M., Rueckert, D., Wachinger, C., Wiestler, B., 2024. Mamba? catch the hype or rethink what really helps for image registration, in: *International Workshop on Biomedical Image Registration*, Springer. pp. 86–97.
- Jian, B., Pan, J., Li, Y., Bongratz, F., Li, R., Rueckert, D., Wiestler, B., Wachinger, C., 2025. Timeflow: Longitudinal brain image registration and aging progression analysis. *arXiv preprint arXiv:2501.08667*.
- Kang, M., Hu, X., Huang, W., et al., 2022. Dual-stream pyramid registration network. *Medical image analysis* 78, 102379.
- Kim, B., Han, I., Ye, J.C., 2022. Diffusemorph: Unsupervised deformable image registration using diffusion model, in: *European conference on computer vision*, Springer. pp. 347–364.
- Kim, B., Kim, D.H., Park, S.H., Kim, J., Lee, J.G., Ye, J.C., 2021. Cyclemorph: cycle consistent unsupervised deformable image registration. *Medical image analysis* 71, 102036.
- Klein, A., Mensh, B., Ghosh, S., Tourville, J., Hirsch, J., 2005. Mindboggle: automated brain labeling with multiple atlases. *BMC medical imaging* 5, 1–14.
- Leow, A.D., Yanovsky, I., Chiang, M.C., Lee, A.D., Klunder, A.D., Lu, A., Becker, J.T., Davis, S.W., Toga, A.W., Thompson, P.M., 2007. Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE transactions on medical imaging* 26, 822–832.
- Li, J., Liu, X., Zhang, F., Cao, X., Buhmann, J.M., Zhang, Y., Li, X., 2025. Gaussian primitives for deformable image registration. *Physics and Imaging in Radiation Oncology*, 100821.
- Liu, H., Ruan, D., Sheng, K., 2025. Unsupervised deformable image registration revisited: Enhancing performance with registration-specific designs, in: *Medical Imaging with Deep Learning-Short Papers*.
- Liu, Y., Chen, J., Wei, S., Carass, A., Prince, J., 2024a. On finite difference jacobian computation in deformable image registration. *International Journal of Computer Vision*, 1–11.
- Liu, Y., Chen, J., Zuo, L., Carass, A., Prince, J.L., 2024b. Vector field attention for deformable image registration. *Journal of Medical Imaging* 11, 064001–064001.
- Liu, Z., Lin, Y., Cao, Y., et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022.
- Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 91–110.
- Lv, J., Wang, Z., Shi, H., Zhang, H., Wang, S., Wang, Y., Li, Q., 2022. Joint progressive and coarse-to-fine registration of brain mri via deformation field integration and non-rigid feature fusion. *IEEE Transactions on Medical Imaging* 41, 2788–2802.
- Ma, J., Li, F., Wang, B., 2024a. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*.
- Ma, T., Dai, X., Zhang, S., Wen, Y., 2023. Pivit: Large deformation image registration with pyramid-iterative vision transformer, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 602–612.
- Ma, T., Zhang, S., Li, J., Wen, Y., 2024b. Iirp-net: Iterative inference residual pyramid network for enhanced image registration, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11546–11555.
- Makela, T., Clarysse, P., Sipila, O., Pauna, N., Pham, Q.C., Katila, T., Magnin, I.E., 2002. A review of cardiac image registration methods. *IEEE Transactions on medical imaging* 21, 1011–1021.
- Marcus, D.S., Wang, T.H., Parker, J., et al., 2007. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience* 19, 1498–1507.
- Meng, M., Bi, L., Feng, D., Kim, J., 2022. Non-iterative coarse-to-fine registration based on single-pass deep cumulative learning, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer. pp. 88–97.
- Meng, M., Feng, D., Bi, L., Kim, J., 2024. Correlation-aware coarse-to-fine mpls for deformable medical image registration, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9645–9654.
- Miao, S., Piat, S., Fischer, P., Tuysuzoglu, A., Mewes, P., Mansi, T., Liao, R., 2018. Dilated fcn for multi-agent 2d/3d medical image registration, in: *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Mok, T.C., Chung, A., 2020a. Fast symmetric diffeomorphic image registration with convolutional neural networks, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4644–4653.
- Mok, T.C., Chung, A., 2022. Affine medical image registration with coarse-to-fine vision transformer, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20835–20844.
- Mok, T.C., Chung, A.C., 2020b. Large deformation diffeomorphic image registration with laplacian pyramid networks, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III* 23, Springer. pp. 211–221.
- Murphy, K., Van Ginneken, B., Reinhardt, J.M., Kabus, S., Ding, K., Deng, X., Cao, K., Du, K., Christensen, G.E., Garcia, V., et al., 2011. Evaluation of registration methods on thoracic ct: the empire10 challenge. *IEEE transactions on medical imaging* 30, 1901–1920.
- Pan, J., Hamdi, M., Huang, W., Hammernik, K., Kuestner, T., Rueckert, D., 2024. Unrolled and rapid motion-compensated reconstruction for cardiac cine mri. *Medical Image Analysis* 91, 103017.
- Pan, J., Rueckert, D., Küstner, T., Hammernik, K., 2021. Efficient image regis-

- tration network for non-rigid cardiac motion estimation, in: Machine Learning for Medical Image Reconstruction, MICCAI 2021, pp. 14–24.
- Pluim, J.P., Maintz, J.A., Viergever, M.A., 2003. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging* 22, 986–1004.
- Qin, Y., Li, X., 2023. Fsdiffreg: Feature-wise and score-wise diffusion-guided unsupervised deformable image registration for cardiac images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 655–665.
- Qiu, H., Qin, C., Schuh, A., Hammernik, K., Rueckert, D., . Learning diffeomorphic and modality-invariant registration using b-splines, in: Medical imaging with deep learning.
- Raschka, S., 2025. The big llm architecture comparison – ahead of ai. URL: <https://magazine.sebastianraschka.com/p/the-big-llm-architecture-comparison>. accessed: 2025-10-01.
- Reinhardt, J.M., Ding, K., Cao, K., Christensen, G.E., Hoffman, E.A., Bodas, S.V., 2008. Registration-based estimates of local lung tissue expansion compared to xenon ct measures of specific ventilation. *Medical image analysis* 12, 752–763.
- Reithmeir, A., Spieker, V., Sideri-Lampretsa, V., Rueckert, D., Schnabel, J.A., Zimmer, V.A., 2025. From model based to learned regularization in medical image registration: A comprehensive review. *Medical Image Analysis* , 103854.
- Risholm, P., Golby, A.J., Wells, W., 2011. Multimodal image registration for preoperative planning and image-guided neurosurgical procedures. *Neurosurgery Clinics* 22, 197–206.
- Rohlfing, T., 2011. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. *IEEE transactions on medical imaging* 31, 153–163.
- Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 234–241.
- Rueckert, D., Aljabar, P., 1999. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging* 18, 712–721.
- Shattuck, D.W., Mirza, M., Adisetiyo, V., et al., 2008. Construction of a 3d probabilistic atlas of human cortical structures. *Neuroimage* 39, 1064–1080.
- Shi, J., He, Y., Kong, Y., et al., 2022. Xmorpher: Full transformer for deformable medical image registration via cross attention, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 217–226.
- Sideri-Lampretsa, V., McGinnis, J., Qiu, H., Paschali, M., Simson, W., Rueckert, D., 2024. Sinr: Spline-enhanced implicit neural representation for multimodal registration. *Medical Imaging with Deep Learning* , 1462–1474.
- Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G., 2020. Implicit neural representations with periodic activation functions. *Advances in neural information processing systems* 33, 7462–7473.
- Song, X., Chao, H., Xu, X., Guo, H., Xu, S., Turkbey, B., Wood, B.J., Sanford, T., Wang, G., Yan, P., 2022. Cross-modal attention for multi-modal image registration. *Medical Image Analysis* 82, 102612.
- Sun, D., Yang, X., Liu, M.Y., Kautz, J., 2018. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8934–8943.
- Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow, in: Computer Vision–ECCV 2020, Springer. pp. 402–419.
- Tian, L., Greer, H., Vialard, F.X., Kwitt, R., Estépar, R.S.J., Rushmore, R.J., Makris, N., Bouix, S., Niethammer, M., 2023. Gradicon: Approximate diffeomorphisms via gradient inverse consistency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18084–18094.
- Tian, X., Wang, J., Zhang, Y., Chen, X., Hu, R., Li, G., Liu, M., Zhang, H., 2025. Gaussian primitive optimized deformable retinal image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 218–228.
- Tolstikhin, I.O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al., 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems* 34, 24261–24272.
- Van Harten, L.D., Stoker, J., Išgum, I., 2023. Robust deformable image registration using cycle-consistent implicit representations. *IEEE transactions on medical imaging* 43, 784–793.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems* 30.
- Vercauteren, T., Pennec, X., Perchant, A., Ayache, N., 2009. Diffeomorphic demons: Efficient non-parametric image registration. *NeuroImage* 45, S61–S72.
- Wang, A.Q., Evan, M.Y., Dalca, A.V., Sabuncu, M.R., 2023a. A robust and interpretable deep learning framework for multi-modal registration via keypoints. *Medical Image Analysis* 90, 102962.
- Wang, H., Ni, D., Wang, Y., 2023b. Modet: Learning deformable image registration via motion decomposition transformer, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 740–749.
- Wang, H., Ni, D., Wang, Y., 2024. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Transactions on Medical Imaging* 43, 2229–2240.
- Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., et al., 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence* 5, e230024.
- Wen, K., Xie, B., Duan, B., Yan, Y., 2024. Mambareg: Mamba-based disentangled convolutional sparse coding for unsupervised deformable multi-modal image registration. *arXiv preprint arXiv:2411.01399* .
- Wolterink, J.M., Zwienenberg, J.C., Brune, C., 2022. Implicit neural representations for deformable image registration, in: International Conference on medical imaging with deep learning, PMLR. pp. 1349–1359.
- Wu, Y., Jiahao, T.Z., Wang, J., Yushkevich, P.A., Hsieh, M.A., Gee, J.C., 2022. Nodoo: A neural ordinary differential equation based optimization framework for deformable image registration, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 20804–20813.
- Xu, A., Shen, S., Chen, W., Zhang, X., 2025. Lightweight hybrid mamba2 for unsupervised medical image registration. *Medical Physics* 52, e18104.
- Xu, J., Chen, E.Z., Chen, X., Chen, T., Sun, S., 2021. Multi-scale neural odes for 3d medical image registration, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 213–223.
- Xu, Z., Lee, C.P., Heinrich, M.P., Modat, M., Rueckert, D., Ourselin, S., Abramson, R.G., Landman, B.A., 2016. Evaluation of six registration methods for the human abdomen on clinically acquired ct. *IEEE Transactions on Biomedical Engineering* 63, 1563–1572.
- Xue, Z., He, K., Xu, D., Gong, J., 2025. Medical image registration via spatial feature extraction mamba and substrate iterative refinement. *IET Image Processing* 19, e70117.
- Yin, W., Sonke, J.J., Gavves, E., 2023. Pc-reg: A pyramidal prediction–correction approach for large deformation image registration. *Medical image analysis* 90, 102978.
- Yu, F., Koltun, V., 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* .
- Zhao, S., Dong, Y., Chang, E.I., Xu, Y., et al., 2019. Recursive cascaded networks for unsupervised medical image registration, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10600–10610.
- Zheng, J.Q., Wang, Z., Huang, B., Lim, N.H., Papież, B.W., 2024. Residual aligner-based network (ran): Motion-separable structure for coarse-to-fine discontinuous deformable registration. *Medical Image Analysis* 91, 103038.

Appendix A. Appendix

Appendix A.1. Progressive Refinement Analysis

Table A.7: Progressive refinement analysis on LPBA (40 subjects), and MindBoggle (100 subjects). OASIS (84 subjects), ADNI (43 subjects), and IXI (115 subjects) datasets, using 200 randomly sampled image pairs each. Reported metrics include Dice score (DSC, %), and deformation regularity measures: SDlogJ, ($\times 10^{-2}$). Metrics are reported across pyramid levels (1/16 $\rightarrow$ 1/2 resolution) and refinement iterations (Iter. 1, Iter. 2).

		LPBA		MindBoggle		OASIS		ADNI		IXI		
	Res.	Iter.	DSC $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	SDlogJ $\downarrow$
initial			54.0 $\pm$ 14.2	-	51.5 $\pm$ 24.5	-	57.3 $\pm$ 22.4	-	52.8 $\pm$ 22.1	-	54.5 $\pm$ 23.7	-
DWP	1/16	-	64.3 $\pm$ 10.8	2.4 $\pm$ 0.5	56.8 $\pm$ 24.7	1.8 $\pm$ 0.2	63.7 $\pm$ 22.0	2.0 $\pm$ 0.3	61.2 $\pm$ 21.4	2.0 $\pm$ 0.2	60.4 $\pm$ 23.7	2.1 $\pm$ 0.3
	1/8	-	68.1 $\pm$ 10.3	3.8 $\pm$ 0.7	61.4 $\pm$ 24.4	3.7 $\pm$ 0.3	71.0 $\pm$ 20.3	3.8 $\pm$ 0.4	67.5 $\pm$ 20.0	3.8 $\pm$ 0.3	66.4 $\pm$ 22.7	3.8 $\pm$ 0.4
	1/4	-	70.0 $\pm$ 10.5	4.4 $\pm$ 0.6	64.9 $\pm$ 24.3	4.9 $\pm$ 0.3	76.0 $\pm$ 19.2	5.1 $\pm$ 0.5	71.9 $\pm$ 19.2	5.0 $\pm$ 0.3	70.1 $\pm$ 22.0	5.2 $\pm$ 0.4
	1/2	-	70.8 $\pm$ 10.7	6.3 $\pm$ 0.6	67.7 $\pm$ 24.1	7.5 $\pm$ 0.4	80.0 $\pm$ 17.6	7.7 $\pm$ 0.5	75.2 $\pm$ 18.3	7.7 $\pm$ 0.4	72.4 $\pm$ 21.1	8.0 $\pm$ 0.6
DWCP	1/16	-	64.1 $\pm$ 10.8	2.6 $\pm$ 0.9	55.9 $\pm$ 24.8	1.9 $\pm$ 0.2	63.7 $\pm$ 21.7	1.9 $\pm$ 0.3	60.8 $\pm$ 21.4	2.0 $\pm$ 0.3	59.9 $\pm$ 23.6	2.1 $\pm$ 0.4
	1/8	-	68.2 $\pm$ 9.8	3.3 $\pm$ 0.7	60.4 $\pm$ 24.7	3.5 $\pm$ 0.3	71.5 $\pm$ 20.2	3.6 $\pm$ 0.5	67.2 $\pm$ 20.0	3.5 $\pm$ 0.4	66.3 $\pm$ 22.7	3.7 $\pm$ 0.4
	1/4	-	70.5 $\pm$ 9.9	3.9 $\pm$ 0.5	64.6 $\pm$ 24.4	4.8 $\pm$ 0.3	77.1 $\pm$ 18.9	5.0 $\pm$ 0.4	72.2 $\pm$ 19.3	4.9 $\pm$ 0.3	70.9 $\pm$ 21.8	5.2 $\pm$ 0.4
	1/2	-	71.6 $\pm$ 10.0	5.3 $\pm$ 1.2	67.6 $\pm$ 24.0	6.9 $\pm$ 0.4	80.9 $\pm$ 17.0	9.6 $\pm$ 2.6	75.5 $\pm$ 18.1	7.2 $\pm$ 0.4	73.1 $\pm$ 20.9	7.6 $\pm$ 0.8
DWCPI	1/16	-	64.0 $\pm$ 10.8	2.6 $\pm$ 0.8	55.5 $\pm$ 24.8	1.8 $\pm$ 0.2	63.3 $\pm$ 22.0	1.8 $\pm$ 0.3	60.8 $\pm$ 21.1	2.0 $\pm$ 0.3	59.9 $\pm$ 23.7	2.0 $\pm$ 0.4
	1/8	-	67.1 $\pm$ 10.1	3.3 $\pm$ 0.7	58.9 $\pm$ 24.9	3.5 $\pm$ 0.4	70.3 $\pm$ 20.8	3.5 $\pm$ 0.5	66.3 $\pm$ 20.4	3.5 $\pm$ 0.4	65.2 $\pm$ 23.1	3.5 $\pm$ 0.4
	1/4	1	69.4 $\pm$ 10.1	3.5 $\pm$ 0.6	63.2 $\pm$ 24.6	4.0 $\pm$ 0.3	75.7 $\pm$ 19.4	4.1 $\pm$ 0.5	70.9 $\pm$ 19.5	4.1 $\pm$ 0.3	69.7 $\pm$ 22.2	4.2 $\pm$ 0.4
	1/4	2	70.0 $\pm$ 10.1	4.0 $\pm$ 0.6	64.2 $\pm$ 24.4	5.0 $\pm$ 0.3	77.0 $\pm$ 19.0	5.0 $\pm$ 0.5	72.1 $\pm$ 19.2	5.0 $\pm$ 0.3	70.6 $\pm$ 21.9	5.2 $\pm$ 0.4
	1/2	1	71.2 $\pm$ 10.3	4.7 $\pm$ 0.6	66.8 $\pm$ 24.1	6.1 $\pm$ 0.3	80.3 $\pm$ 17.4	6.7 $\pm$ 0.9	74.8 $\pm$ 18.4	6.2 $\pm$ 0.3	72.7 $\pm$ 21.1	6.4 $\pm$ 0.4
	1/2	2	71.3 $\pm$ 10.3	5.6 $\pm$ 0.6	67.7 $\pm$ 23.9	7.2 $\pm$ 0.4	81.1 $\pm$ 16.8	10.2 $\pm$ 3.1	75.6 $\pm$ 18.0	7.4 $\pm$ 0.4	73.2 $\pm$ 20.7	7.7 $\pm$ 0.5

Table A.8: Progressive refinement analysis on NLST (in-domain) and Lung250M-4B (zero-shot) datasets. Reported metrics include target registration error (TRE, mm or vox), and deformation regularity measures: SDlogJ, ($\times 10^{-2}$). Metrics are reported across pyramid levels (1/16 $\rightarrow$ 1/2 resolution) and refinement iterations (Iter. 1, Iter. 2).

			Lung250M-4B				NLST			
		Iter.	INP $\rightarrow$ EXP		EXP $\rightarrow$ INP		INP $\rightarrow$ EXP		EXP $\rightarrow$ INP	
Res.			TRE $\downarrow$	SDlogJ $\downarrow$	TRE $\downarrow$	SDlogJ $\downarrow$	TRE $\downarrow$	SDlogJ $\downarrow$	TRE $\downarrow$	SDlogJ $\downarrow$
initial			13.74 $\pm$ 3.63	-	13.74 $\pm$ 3.63	-	10.14 $\pm$ 2.92	-	10.14 $\pm$ 2.92	-
DWP	1/16	-	11.49 $\pm$ 3.42	5.7 $\pm$ 1.7	10.82 $\pm$ 3.39	5.2 $\pm$ 1.4	5.77 $\pm$ 1.72	4.1 $\pm$ 1.0	5.64 $\pm$ 1.80	4.1 $\pm$ 1.0
	1/8	-	10.16 $\pm$ 3.84	8.0 $\pm$ 1.7	9.89 $\pm$ 3.78	6.8 $\pm$ 1.4	3.96 $\pm$ 1.46	5.3 $\pm$ 1.1	3.91 $\pm$ 1.47	5.5 $\pm$ 1.2
	1/4	-	9.78 $\pm$ 4.14	10.5 $\pm$ 2.2	9.53 $\pm$ 4.06	8.7 $\pm$ 1.7	3.01 $\pm$ 1.41	6.6 $\pm$ 1.3	3.00 $\pm$ 1.40	6.7 $\pm$ 1.4
	1/2	-	9.64 $\pm$ 4.22	13.1 $\pm$ 2.7	9.51 $\pm$ 4.16	10.9 $\pm$ 2.1	2.60 $\pm$ 1.44	7.9 $\pm$ 1.5	2.60 $\pm$ 1.41	7.9 $\pm$ 1.5
DWCP	1/16	-	10.38 $\pm$ 3.11	6.2 $\pm$ 1.1	9.92 $\pm$ 3.27	4.7 $\pm$ 0.6	5.34 $\pm$ 1.90	4.0 $\pm$ 0.9	5.28 $\pm$ 1.81	4.1 $\pm$ 0.9
	1/8	-	8.94 $\pm$ 3.46	8.6 $\pm$ 1.9	8.34 $\pm$ 3.57	6.4 $\pm$ 1.0	3.57 $\pm$ 1.55	5.5 $\pm$ 1.1	3.56 $\pm$ 1.55	5.6 $\pm$ 1.2
	1/4	-	8.26 $\pm$ 3.86	10.6 $\pm$ 2.2	7.77 $\pm$ 3.83	7.8 $\pm$ 1.1	2.43 $\pm$ 1.47	6.6 $\pm$ 1.2	2.44 $\pm$ 1.47	6.7 $\pm$ 1.3
	1/2	-	8.04 $\pm$ 4.02	12.7 $\pm$ 3.1	7.58 $\pm$ 3.91	9.5 $\pm$ 1.7	2.00 $\pm$ 1.45	7.5 $\pm$ 1.2	2.03 $\pm$ 1.46	7.6 $\pm$ 1.5
DWCPI	1/16	-	10.55 $\pm$ 3.28	6.3 $\pm$ 1.3	10.16 $\pm$ 3.45	5.1 $\pm$ 0.7	5.45 $\pm$ 1.87	4.0 $\pm$ 0.8	5.45 $\pm$ 1.90	4.1 $\pm$ 0.9
	1/8	-	9.51 $\pm$ 3.58	8.8 $\pm$ 2.0	8.87 $\pm$ 3.78	6.5 $\pm$ 1.0	4.03 $\pm$ 1.60	5.4 $\pm$ 1.0	4.03 $\pm$ 1.60	5.5 $\pm$ 1.1
	1/4	1	8.99 $\pm$ 4.02	10.1 $\pm$ 2.4	8.36 $\pm$ 4.13	7.2 $\pm$ 1.2	2.85 $\pm$ 1.57	6.0 $\pm$ 1.1	2.88 $\pm$ 1.57	6.2 $\pm$ 1.3
	1/4	2	8.73 $\pm$ 4.18	11.0 $\pm$ 2.4	8.09 $\pm$ 4.24	8.0 $\pm$ 1.2	2.50 $\pm$ 1.49	6.8 $\pm$ 1.1	2.54 $\pm$ 1.50	6.9 $\pm$ 1.2
	1/2	1	8.52 $\pm$ 4.33	11.8 $\pm$ 2.5	7.94 $\pm$ 4.30	8.9 $\pm$ 1.5	2.06 $\pm$ 1.49	7.1 $\pm$ 1.2	2.10 $\pm$ 1.50	7.2 $\pm$ 1.3
	1/2	2	8.42 $\pm$ 4.37	12.6 $\pm$ 2.7	7.88 $\pm$ 4.31	9.6 $\pm$ 1.4	1.95 $\pm$ 1.47	7.6 $\pm$ 1.1	2.00 $\pm$ 1.49	7.7 $\pm$ 1.3

Table A.9: Progressive refinement analysis on ACDC and M&Ms (zero-shot) datasets. Reported metrics include Dice score (DSC, %), normalized surface Dice (NSD, %) and deformation regularity measures: SDlogJ, ($\times 10^{-2}$). Metrics are reported across pyramid levels (1/16 $\rightarrow$ 1/2 resolution) and refinement iterations (Iter. 1, Iter. 2).

			M&Ms						ACDC Test							
			ED $\rightarrow$ ES			ES $\rightarrow$ ED			ED $\rightarrow$ ES			ES $\rightarrow$ ED				
			Res.	Iter.	DSC $\uparrow$	NSD $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	NSD $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	NSD $\uparrow$	SDlogJ $\downarrow$	DSC $\uparrow$	NSD $\uparrow$	SDlogJ $\downarrow$
initial					60.5 $\pm$ 12.7	8.0 $\pm$ 6.0	-	60.5 $\pm$ 12.7	8.0 $\pm$ 6.0	-	60.2 $\pm$ 14.8	10.7 $\pm$ 7.3	-	60.2 $\pm$ 14.8	10.7 $\pm$ 7.3	-
DWP	1/16	-	72.7 $\pm$ 12.0	12.6 $\pm$ 5.1	2.0 $\pm$ 0.7	74.3 $\pm$ 11.3	10.3 $\pm$ 5.1	1.8 $\pm$ 0.6	69.5 $\pm$ 14.3	15.2 $\pm$ 5.8	2.3 $\pm$ 0.8	70.7 $\pm$ 13.6	11.8 $\pm$ 5.5	1.8 $\pm$ 0.6		
	1/8	-	77.6 $\pm$ 10.6	15.3 $\pm$ 5.6	2.5 $\pm$ 0.8	81.1 $\pm$ 9.2	14.5 $\pm$ 6.7	2.3 $\pm$ 0.7	76.1 $\pm$ 12.7	18.3 $\pm$ 6.1	3.4 $\pm$ 1.2	79.5 $\pm$ 11.2	16.4 $\pm$ 6.4	3.1 $\pm$ 1.0		
	1/4	-	79.8 $\pm$ 9.7	17.9 $\pm$ 6.0	3.0 $\pm$ 1.0	84.0 $\pm$ 8.1	16.9 $\pm$ 7.2	2.7 $\pm$ 0.8	79.3 $\pm$ 11.5	21.9 $\pm$ 6.0	4.3 $\pm$ 1.5	83.5 $\pm$ 9.5	20.4 $\pm$ 6.9	3.7 $\pm$ 1.2		
	1/2	-	80.3 $\pm$ 9.5	18.8 $\pm$ 6.3	3.5 $\pm$ 1.2	85.3 $\pm$ 7.6	18.8 $\pm$ 7.8	3.2 $\pm$ 1.2	80.3 $\pm$ 10.9	23.3 $\pm$ 5.7	4.9 $\pm$ 1.6	85.4 $\pm$ 8.5	23.2 $\pm$ 7.5	4.3 $\pm$ 1.3		
DWCP	1/16	-	71.1 $\pm$ 12.1	11.6 $\pm$ 4.5	2.1 $\pm$ 0.9	73.6 $\pm$ 11.5	10.1 $\pm$ 4.8	2.0 $\pm$ 0.8	69.9 $\pm$ 13.7	15.3 $\pm$ 5.8	2.4 $\pm$ 0.8	70.2 $\pm$ 13.6	11.6 $\pm$ 5.5	1.8 $\pm$ 0.6		
	1/8	-	77.8 $\pm$ 11.1	15.1 $\pm$ 5.4	2.7 $\pm$ 1.1	81.9 $\pm$ 9.4	15.3 $\pm$ 6.6	2.5 $\pm$ 0.9	76.5 $\pm$ 12.3	18.5 $\pm$ 6.2	3.4 $\pm$ 1.3	80.1 $\pm$ 10.6	16.6 $\pm$ 6.0	3.1 $\pm$ 1.1		
	1/4	-	80.9 $\pm$ 9.8	18.1 $\pm$ 6.0	3.2 $\pm$ 1.2	85.1 $\pm$ 8.0	18.6 $\pm$ 7.1	2.9 $\pm$ 1.2	80.1 $\pm$ 11.1	22.5 $\pm$ 5.8	4.3 $\pm$ 1.6	84.1 $\pm$ 8.9	20.9 $\pm$ 6.8	3.8 $\pm$ 1.2		
	1/2	-	81.3 $\pm$ 9.6	19.0 $\pm$ 6.3	3.9 $\pm$ 3.5	86.2 $\pm$ 7.5	20.3 $\pm$ 7.5	3.5 $\pm$ 3.7	81.3 $\pm$ 10.3	23.9 $\pm$ 5.7	5.2 $\pm$ 3.5	86.1 $\pm$ 7.9	24.4 $\pm$ 7.8	4.4 $\pm$ 1.3		
DWCPI	1/16	-	71.1 $\pm$ 12.1	11.9 $\pm$ 4.7	2.2 $\pm$ 0.9	73.1 $\pm$ 11.6	10.2 $\pm$ 4.7	2.0 $\pm$ 0.8	70.1 $\pm$ 13.7	15.4 $\pm$ 6.0	2.4 $\pm$ 0.8	70.5 $\pm$ 13.6	11.4 $\pm$ 5.4	1.8 $\pm$ 0.6		
	1/8	-	77.9 $\pm$ 11.1	15.2 $\pm$ 5.6	2.7 $\pm$ 1.1	81.5 $\pm$ 9.5	15.3 $\pm$ 6.5	2.5 $\pm$ 0.9	76.2 $\pm$ 12.4	18.2 $\pm$ 6.4	3.4 $\pm$ 1.3	79.5 $\pm$ 11.1	16.4 $\pm$ 6.3	3.1 $\pm$ 1.0		
	1/4	1	80.3 $\pm$ 10.3	17.5 $\pm$ 6.0	2.8 $\pm$ 1.2	84.0 $\pm$ 8.7	17.4 $\pm$ 7.1	2.6 $\pm$ 1.0	78.9 $\pm$ 11.6	21.4 $\pm$ 6.0	3.7 $\pm$ 1.4	82.2 $\pm$ 10.2	18.6 $\pm$ 6.6	3.3 $\pm$ 1.1		
	1/4	2	81.2 $\pm$ 9.8	18.5 $\pm$ 6.2	3.2 $\pm$ 1.4	85.2 $\pm$ 8.1	18.7 $\pm$ 7.2	2.9 $\pm$ 1.3	80.0 $\pm$ 11.1	22.4 $\pm$ 5.9	4.3 $\pm$ 1.6	83.8 $\pm$ 9.6	20.7 $\pm$ 6.9	3.7 $\pm$ 1.2		
	1/2	1	81.6 $\pm$ 9.6	19.0 $\pm$ 6.3	3.8 $\pm$ 4.3	85.8 $\pm$ 7.7	19.5 $\pm$ 7.5	3.5 $\pm$ 4.8	80.9 $\pm$ 10.7	23.8 $\pm$ 5.9	4.6 $\pm$ 1.6	85.0 $\pm$ 8.9	22.3 $\pm$ 7.5	4.0 $\pm$ 1.3		
	1/2	2	81.7 $\pm$ 9.5	19.0 $\pm$ 6.3	4.3 $\pm$ 5.3	86.2 $\pm$ 7.4	20.1 $\pm$ 7.5	3.9 $\pm$ 5.3	81.2 $\pm$ 10.4	23.8 $\pm$ 5.7	5.0 $\pm$ 1.7	85.8 $\pm$ 8.4	23.6 $\pm$ 7.9	4.4 $\pm$ 1.3		

Appendix A.2. Abdomen MRI-CT DSC Breakdown

Table A.10: Dice score for each subject and organ on Abdomen MRI → CT registration task. The best (■, **bold**), second best (■, *italic*), and third best (■) results are highlighted.

	1				2				3				4			
	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid
initial	71.46	57.84	-	55.32	76.59	<i>71.63</i>	-	70.20	67.65	<i>71.04</i>	55.40	74.50	82.44	77.60	75.70	74.91
VXM	80.38	59.49	-	63.60	83.26	67.20	-	67.73	68.68	70.71	66.72	76.48	84.41	64.51	70.27	66.92
Mam-VXM	80.00	56.61	-	60.77	84.82	67.21	-	66.90	<i>67.72</i>	58.65	59.50	72.68	83.40	66.62	<i>77.97</i>	70.24
TM	79.56	<i>60.32</i>	-	64.28	<i>84.47</i>	68.98	-	73.04	67.15	64.39	60.70	<i>76.29</i>	83.31	66.70	68.17	62.00
Mam-TM	78.80	59.66	-	61.93	83.70	70.30	-	<i>70.89</i>	66.94	63.18	63.16	73.68	81.51	70.99	73.02	63.51
LKU	78.65	56.39	-	54.34	79.33	58.46	-	62.78	62.22	45.80	59.69	69.91	<i>87.04</i>	69.79	77.11	72.11
Dual	<i>82.99</i>	56.56	-	<i>69.76</i>	83.03	65.50	-	66.27	<i>68.25</i>	55.44	<i>64.01</i>	71.57	85.67	70.03	70.52	71.85
DWP	85.21	56.78	-	68.47	82.62	73.02	-	48.76	65.72	67.20	<i>66.27</i>	70.48	82.00	<i>78.71</i>	75.73	74.05
DWCP	81.16	64.30	-	77.57	84.41	70.34	-	64.62	66.44	71.86	46.17	54.75	89.33	81.68	<i>84.67</i>	83.82
DWCPI	<i>81.26</i>	<i>64.02</i>	-	<i>76.73</i>	80.43	60.63	-	59.80	64.10	60.79	50.44	57.16	<i>89.00</i>	<i>77.96</i>	86.13	<i>83.57</i>
	5				6				7				8			
	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid
initial	55.52	56.22	51.56	52.58	71.52	60.08	50.06	47.67	45.42	65.26	58.88	57.53	71.93	71.15	<i>72.23</i>	<i>73.30</i>
VXM	65.32	75.51	48.85	54.10	81.18	48.24	<i>72.85</i>	58.82	65.64	72.43	73.21	67.90	76.13	75.85	68.82	71.43
Mam-VXM	68.36	73.90	<i>62.22</i>	65.36	80.69	40.01	65.57	72.39	63.58	68.39	65.33	65.24	79.02	<i>80.08</i>	66.91	63.02
TM	<i>74.69</i>	<i>77.62</i>	46.77	56.98	<i>81.62</i>	48.95	61.62	72.38	65.75	<i>71.50</i>	64.55	75.40	75.92	74.99	63.53	52.45
Mam-TM	<i>73.32</i>	75.11	57.82	63.08	80.32	45.23	64.66	68.29	65.95	<i>72.42</i>	<i>69.22</i>	<i>73.08</i>	<i>80.05</i>	78.09	69.85	67.68
LKU	54.57	56.05	58.44	52.42	72.74	67.40	62.52	53.10	<i>73.98</i>	63.68	58.77	59.72	72.02	71.44	67.98	70.66
Dual	57.77	69.56	47.29	51.22	77.86	38.13	70.23	<i>73.23</i>	65.39	66.62	<i>70.34</i>	59.42	74.34	75.44	71.17	<i>73.23</i>
DWP	67.49	74.45	55.26	<i>67.25</i>	80.30	56.58	68.51	67.24	<i>72.61</i>	67.42	60.95	55.98	79.16	76.23	61.56	63.07
DWCP	81.19	<i>82.35</i>	66.30	78.31	85.11	60.13	73.61	<i>73.58</i>	71.57	70.19	62.82	66.93	<i>81.48</i>	<i>80.63</i>	<i>73.47</i>	71.17
DWCPI	70.80	82.52	<i>62.77</i>	<i>71.95</i>	<i>84.92</i>	<i>61.47</i>	<i>72.08</i>	73.87	79.08	66.37	64.36	<i>70.73</i>	83.16	81.76	76.54	73.47

Table A.11: Dice score for each subject and organ on Abdomen CT → MRI registration task. The best (■, **bold**), second best (□, *italic*), and third best (●) results are highlighted.

	1				2				3				4			
	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid
initial	71.46	57.84	-	55.32	76.59	71.63	-	70.20	67.65	71.04	55.40	74.50	82.44	77.60	75.70	74.91
VXM	78.67	58.59	-	75.10	82.70	57.17	-	75.51	65.58	59.51	58.92	62.61	82.73	70.29	72.62	67.11
Mam-VXM	76.36	57.89	-	67.24	81.80	60.75	-	74.41	65.68	50.98	53.95	66.01	82.44	62.70	71.06	71.73
TM	72.06	60.17	-	67.61	81.20	63.08	-	76.89	63.55	62.37	57.70	70.11	80.77	70.03	68.20	64.80
Mam-TM	74.37	58.12	-	66.88	81.30	59.21	-	79.29	64.89	63.27	56.21	69.51	80.62	69.88	67.57	66.44
LKU	70.84	54.71	-	53.44	81.72	64.47	-	49.67	62.92	57.76	57.80	67.35	82.53	70.28	81.38	80.73
Dual	77.80	53.61	-	72.00	82.29	58.98	-	74.25	66.53	59.39	53.72	62.17	83.16	71.68	64.44	73.44
DWP	79.47	49.98	-	69.44	83.78	64.64	-	70.21	66.21	62.42	59.37	69.88	82.26	78.10	76.37	78.76
DWCP	79.53	64.83	-	77.84	82.32	64.47	-	74.44	67.33	69.08	43.88	70.38	89.80	82.30	84.38	88.10
DWCPI	76.67	65.20	-	76.21	81.09	67.84	-	68.49	62.26	59.92	52.16	67.45	89.25	82.51	84.54	86.90
	5				6				7				8			
	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid	Liver	Spleen	L-Kid	R-Kid
initial	55.52	56.22	51.56	52.58	71.52	60.08	50.06	47.67	45.42	65.26	58.88	57.53	71.93	71.15	72.23	73.30
VXM	66.92	71.03	54.17	62.47	77.57	44.32	65.13	69.62	62.82	73.95	78.50	75.29	72.38	74.24	69.58	76.00
Mam-VXM	65.14	70.15	61.78	62.60	78.18	48.60	68.17	64.78	60.77	74.11	69.58	69.13	75.62	73.78	68.08	68.85
TM	67.39	74.74	57.60	60.96	74.93	46.43	58.67	69.07	60.74	76.45	70.79	74.98	73.28	81.71	63.56	64.05
Mam-TM	64.96	70.50	52.37	62.80	72.12	47.92	61.20	67.17	63.60	74.73	70.51	76.57	76.62	81.17	60.59	59.14
LKU	58.53	52.75	53.93	52.71	68.14	51.27	54.60	39.77	63.59	64.19	59.30	60.92	75.32	83.82	70.30	70.37
Dual	60.77	67.76	47.71	52.25	77.54	40.43	73.68	74.81	61.64	69.67	78.89	65.57	71.06	74.11	66.36	78.05
DWP	63.64	69.16	60.28	67.98	80.67	51.31	73.32	65.97	67.48	71.17	69.65	64.06	75.97	77.86	66.68	70.98
DWCP	66.50	80.58	72.76	74.13	82.88	49.25	70.14	54.58	73.46	72.73	66.63	72.93	83.53	84.15	73.01	67.78
DWCPI	69.25	76.55	64.21	71.10	81.59	42.09	69.21	51.94	72.79	74.05	69.94	68.00	82.37	83.58	72.42	72.58