LEARNING RELIABLE EVIDENCE FOR COARSE-TO-FINE DEFORMABLE MEDICAL IMAGE REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical image registration is fundamental to quantitative anatomical analysis and the assessment of pathological changes. Existing methods commonly use multi-scale features to progressively align global structures and local details. However, downsampling can introduce aliasing into coarse-scale representations when high-frequency components are insufficiently attenuated. The resulting changes in local feature responses can disrupt anatomical correspondence estimation. We propose EviGuardMorph, a framework that improves registration reliability through the construction and use of correspondence evidence. The Frequency-Biased Evidence Encoder (FBE) applies scale-dependent Gaussian prefiltering before downsampling to construct structure and detail evidence streams with different effective bandwidths. The detail stream is retained only at high-resolution scales, where it interacts with the structure stream through reversible coupling. The Baseline-Protected Bidirectional Matcher (BPBM) preserves the one-way local matching estimate as a baseline. It combines matching entropy with forward-backward cycle consistency to weight bidirectional candidate residuals, suppressing erroneous updates from ambiguous matches. Experiments demonstrate that EviGuardMorph achieves leading registration accuracy across multiple datasets. Spectral analysis further shows that prefiltering reduces the out-of-band spectral energy fraction before downsampling by 90.4%–99.7%.

1 INTRODUCTION

Medical image registration estimates underlying spatial transformations to establish anatomical correspondences across images from different individuals, time points, or imaging modalities, providing a foundation for quantitative anatomical analysis and disease progression assessment (Balakrishnan et al., 2019). Inter-subject variability, organ motion, and pathological changes introduce global anatomical discrepancies and local non-rigid deformations, requiring accurate registration to reconcile global structural alignment with precise correspondence at organ boundaries and within fine-scale tissues (Cheng et al., 2025; Abulnaga et al., 2026).

Traditional deformable registration methods estimate spatial transformations by iteratively optimizing image similarity, with regularization promoting smooth, physically plausible deformations (Rueckert et al., 1999). Solving each image pair independently, however, incurs substantial computational cost. Learning-based methods, exemplified by VoxelMorph and recent frameworks, learn mappings from image pairs to deformation fields, enabling efficient registration in a single forward pass (Balakrishnan et al., 2019; Abulnaga et al., 2026). To reconcile large-scale deformation with fine-scale alignment, multiscale methods estimate coarse deformations at low resolution and progressively refine local correspondences through hierarchical pre-alignment and residual updates (Cheng et al., 2025). Within this coarse-to-fine framework, explicit feature-matching methods convert local feature similarities into displacement estimates (Wang et al., 2023). Multiscale features therefore serve as both image representations and matching evidence for anatomical correspondence. Their reliability directly governs the accuracy of the resulting deformation estimate.

However, enhanced spatial representations alone do not guarantee reliable correspondence evidence under downsampling. Downsampling lowers the Nyquist limit, causing insufficiently attenuated out-of-band components to alias into coarse-scale representations (Rahman & Yeh, 2025). This can distort local similarity distributions and induce false matches. Fig. 1(b) shows substantial responses to

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

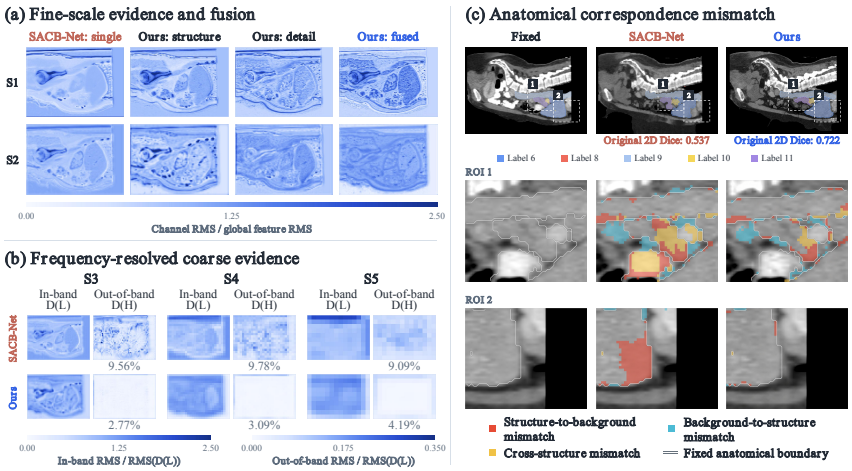


Figure 1: **Motivation for reliable multiscale correspondence evidence.** (a) Fine-scale responses at S1–S2 from SACB-Net and our structure, detail, and fused streams. (b) In-band $D(L)$ and out-of-band $D(H)$ responses at S3–S5. Percentages denote volumetric out-of-band-to-in-band response norm ratios. (c) Anatomical mismatch schematic with two enlarged ROIs. Red, cyan, and yellow indicate structure-to-background, background-to-structure, and cross-structure mismatches, respectively. White contours mark fixed anatomical boundaries.

out-of-band components at S3–S5 in a representative multiscale registration method. Since coarse-scale deformations pre-align finer-scale moving features (Wang et al., 2023; Cheng et al., 2025), early matching errors can propagate through the hierarchy and compromise later refinement. Although low-pass prefiltering alleviates aliasing, uniform smoothing may weaken boundaries and texture required for precise localization (Luo et al., 2026). Similar feature responses at candidate locations can also leave local matching ambiguous even after filtering. Fig. 1(c) illustrates different types of anatomical mismatch. Reliable registration therefore requires scale-adapted evidence that suppresses out-of-band components while preserving fine-scale cues, together with further refinement of displacement estimates guided by evidence reliability.

In this work, we propose EviGuardMorph to improve the reliability of multiscale deformation estimation through anatomical correspondence evidence. The Frequency-Biased Evidence Encoder (FBE) applies scale-dependent Gaussian prefiltering before downsampling to construct structure and detail streams with different effective bandwidths. The structure stream spans the pyramid to mitigate out-of-band interference, while the detail stream is confined to S1–S2, where reversible coupling and subsequent fusion preserve fine-scale localization cues (Fig. 1(a)). The Baseline-Protected Bidirectional Matcher (BPBM) uses matching entropy and forward-backward cycle consistency to weight bidirectional residual corrections relative to the one-way baseline, limiting the influence of ambiguous candidates. Our contributions are summarized as follows:

- We propose EviGuardMorph, a reliable evidence learning framework that improves deformable registration through scale-adapted encoding and reliability-guided matching.
- We develop FBE, combining scale-dependent Gaussian prefiltering with reversible structure-detail coupling to suppress out-of-band components while retaining local detail.
- We introduce BPBM, which uses matching entropy and forward-backward cycle consistency to apply bounded, reliability-weighted bidirectional residual corrections.
- We achieve state-of-the-art registration accuracy on four datasets. Prefiltering yields a 90.4%–99.7% relative reduction in the out-of-band energy fraction before downsampling.

2 Related Work

2.1 Medical Image Registration

Learning-based deformation prediction improves registration efficiency (Balakrishnan et al., 2019), while recent computational frameworks enable large-scale volumetric processing (Jena et al., 2026).

Recent efforts strengthen anatomical representations through spatially adaptive encoding (Cheng et al., 2025), self-supervised pretraining (Fu et al., 2026), and joint surface–volume modeling (Abulnaga et al., 2026) to improve correspondence estimation and anatomical consistency. A complementary line adopts multiscale representations, as exemplified by GroupMorph, ModeT, and RDP (Tan et al., 2024; Wang et al., 2023; 2024), to accommodate large displacements by establishing coarse correspondences and progressively refining them with finer-scale features. Feature-matching approaches extend this paradigm by converting local similarities between hierarchical features into explicit coordinate correspondences that guide successive deformation updates (Liu et al., 2022; Wang et al., 2023). By coupling coarse structural alignment with fine-grained localization, multiscale representations have thus become a cornerstone of modern deformable registration.

2.2 Anti-Aliasing in Multiscale Representations

Multiscale registration requires global structure for reliable coarse-scale correspondence and local detail for precise fine-scale localization (Wang et al., 2023; Cheng et al., 2025). Since downsampling intrinsically lowers the Nyquist limit, high-frequency components essential for high-resolution localization may exceed this reduced threshold. If insufficiently attenuated, these components induce aliasing that perturbs correspondence estimation. Prior work mitigates aliasing through fixed or content-adaptive low-pass filtering before downsampling (Chen et al., 2024; Rahman & Yeh, 2025; Luo et al., 2026). For registration encoders, however, filtering must be coordinated with cross-resolution evidence transfer to accommodate distinct coarse- and fine-scale matching requirements. We combine scale-dependent prefiltering with fine-scale structure–detail coupling, while correspondence reliability controls bidirectional residual corrections.

3 Methods

3.1 Problem Formulation and Overview

Given fixed and moving images I_f and I_m on a three-dimensional grid $\Omega \subset \mathbb{R}^3$, we seek a dense displacement field aligning I_m with I_f. Backward warping is defined as

$$\mathcal{W}(X, u)(x) = X(x + u(x)), \tag{1}$$

where X denotes the input to be warped, x is a spatial location, and u is the sampling displacement field. Non-integer sampling locations are evaluated by trilinear interpolation. The induced spatial mapping is $\mathrm{Id} + u$, with Id denoting the identity map.

As shown in Fig. 2, a shared FBE maps both images to evidence pyramids $\{E_f^s, E_m^s\}_{s=1}^5$. S1 retains the input resolution, and scale s has grid spacing ρ^{s-1} times that of the input, where ρ is the downsampling factor between adjacent scales. BPBM estimates displacement increments from S5 to S2, followed by a Smooth Residual Refinement Head (SRRH) at full resolution (S1).

Hierarchical pre-alignment and displacement estimation. At scale s, $\Delta\varphi_s$ and φ_s denote the estimated displacement increment and accumulated displacement, respectively, in current-scale voxel units. At S5, we set $\widetilde{E}_m^5 = E_m^5$ and $\varphi_5 = \Delta\varphi_5$. At each finer scale $s = 4, 3, 2, 1$, the upsampled displacement pre-aligns the moving evidence. We then estimate $\Delta\varphi_s$ and update the accumulated displacement by composition:

$$\begin{aligned} \widetilde{E}_m^s &= \mathcal{W}(E_m^s, \mathcal{U}(\varphi_{s+1}))\,, \\ \varphi_s &= \Delta\varphi_s + \mathcal{W}(\mathcal{U}(\varphi_{s+1}), \Delta\varphi_s)\,, \end{aligned} \tag{2}$$

where $\mathcal{U}(u) = \rho\ \mathrm{Interp}_\rho(u)$ upsamples trilinearly by ρ and rescales to current-scale voxel units. The composition samples the propagated displacement at locations shifted by $\Delta\varphi_s$.

Full-resolution refinement. At S1, SRRH uses the full-resolution fixed evidence E_f^1 and pre-aligned moving evidence $\widetilde{E}_m^1$ to estimate the final displacement increment. Their channel-wise concatenation is fed to a residual prediction network $\mathcal{H}$, whose three output channels correspond to the three spatial displacement components. A fixed Gaussian kernel smooths each component:

$$\Delta\varphi_1 = G_{\mathrm{ref}} *_{\mathrm{dw}} \mathcal{H}\Big([E_f^1, \widetilde{E}_m^1]\Big)\,, \tag{3}$$

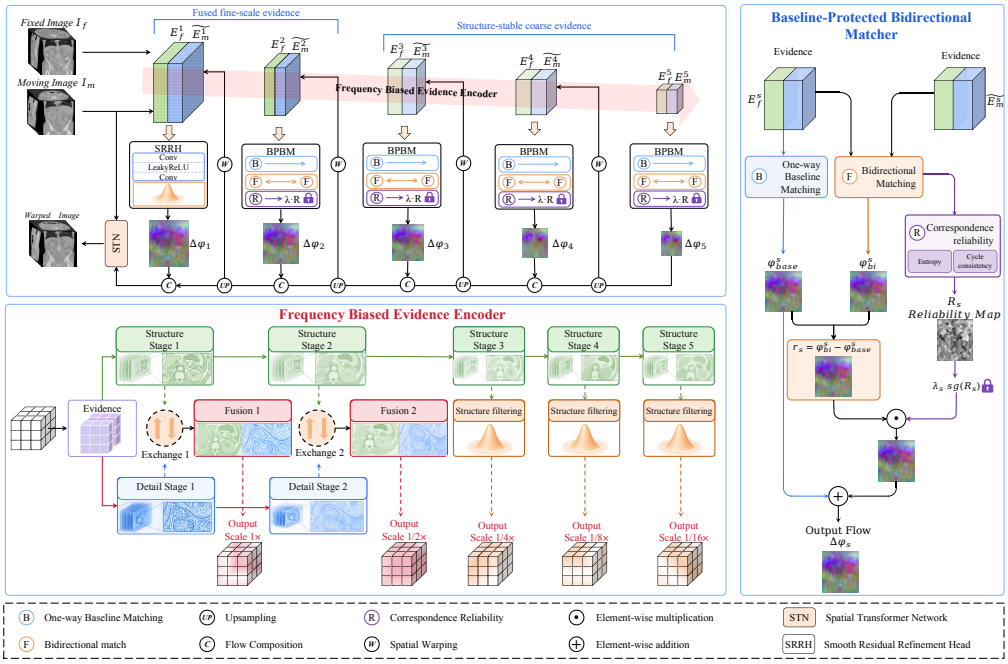


Figure 2: **Overview of EviGuardMorph.** A shared Frequency-Biased Evidence Encoder (FBE) produces fused structure–detail evidence at S1–S2 and structure-stable evidence at S3–S5. Baseline-Protected Bidirectional Matchers (BPBMs) estimate coarse-to-fine updates at S5–S2 by adding reliability-weighted bidirectional residual corrections to the one-way baseline. The propagated deformation pre-aligns finer-scale moving evidence, and SRRH predicts the residual at S1.

where $[\cdot,\cdot]$ denotes channel concatenation and $*_{\mathrm{dw}}$ denotes depthwise convolution. The Gaussian kernel G_{ref} has standard deviation σ_{ref} and width k_{ref}. The smoothed increment $\Delta\varphi_1$ is then composed with the upsampled displacement $\mathcal{U}(\varphi_2)$ according to Eq. (2), yielding the final accumulated displacement φ_1. Finally, φ_1 warps the original moving image to obtain $\widehat{I}_m = \mathcal{W}(I_m, \varphi_1)$.

3.2 Frequency-Biased Evidence Encoder

FBE constructs scale-adapted correspondence evidence for coarse structural matching and fine-scale localization. As shown in Fig. 2, its structure stream spans S1–S5, while the detail stream is restricted to S1–S2. Both images share weights. Fixed and moving subscripts are omitted below.

Scale-dependent prefiltering. We attenuate out-of-band components using fixed depthwise Gaussian filtering followed by average pooling:

$$\mathcal{A}_{\sigma,k}(X) = \mathrm{AvgPool}_{\rho}\big(G_{\sigma,k} *_{\mathrm{dw}} X\big),$$
$$G_{\sigma,k}(q) = \frac{\exp\big(-\|q\|_2^2/(2\sigma^2)\big)}{\sum_{\delta\in\mathcal{B}_k}\exp\big(-\|\delta\|_2^2/(2\sigma^2)\big)}, \tag{4}$$

where ρ is the downsampling factor, k is the odd kernel width, and $q \in \mathcal{B}_k = \{-(k-1)/2, \ldots, (k-1)/2\}^3$. The standard deviation $\sigma > 0$ is measured in voxel units of the feature grid before downsampling. For each structure transition $s \to s+1$, we use $(\sigma_s^{\mathrm{str}}, k_s^{\mathrm{str}})$, increasing σ_s^{str} toward coarser scales to strengthen out-of-band attenuation. The detail stream undergoes only S1-to-S2 downsampling, using $(\sigma_1^{\mathrm{det}}, k_1^{\mathrm{det}})$ with $\sigma_1^{\mathrm{det}} < \sigma_1^{\mathrm{str}}$ to preserve richer local responses.

Structure and detail pathways. A shared convolutional stem feeds independent structure and detail blocks at S1. Each block comprises three-dimensional convolutions, instance normalization, and nonlinear activations. At S1–S2, the streams exchange information before fusion into matching evidence. Cross-scale propagation uses the branch features rather than the fused outputs. At S1,

each branch is prefiltered, downsampled, and processed by a convolutional block to form its S2 input. Only the structure stream continues beyond S2.

Fine-scale reversible coupling. Let $E^s_{\mathrm{str,in}}$ and $E^s_{\mathrm{det,in}}$ denote the outputs of the structure and detail convolutional blocks, respectively, before coupling at scale $s \in \{1, 2\}$. Reversible coupling exchanges information through two learnable transformations (Han et al., 2023):

$$\begin{aligned} \overline{E}^s_{\mathrm{str}} &= E^s_{\mathrm{str,in}} + \mathcal{Q}_s(E^s_{\mathrm{det,in}}), \\ E^s_{\mathrm{det}} &= E^s_{\mathrm{det,in}} + \mathcal{V}_s(\overline{E}^s_{\mathrm{str}}), \end{aligned} \tag{5}$$

where $\mathcal{Q}_s$ and $\mathcal{V}_s$ comprise pointwise convolutions, normalization, and nonlinear activations. Both inputs remain recoverable before output filtering and fusion through $E^s_{\mathrm{det,in}} = E^s_{\mathrm{det}} - \mathcal{V}_s(\overline{E}^s_{\mathrm{str}})$ and $E^s_{\mathrm{str,in}} = \overline{E}^s_{\mathrm{str}} - \mathcal{Q}_s(E^s_{\mathrm{det,in}})$.

Evidence output and cross-scale propagation. At S1–S2, the coupled structure features undergo structure filtering: $E^s_{\mathrm{str}} = \mathcal{R}_s(\overline{E}^s_{\mathrm{str}})$. Here, $\mathcal{R}_1$ is the identity, while $\mathcal{R}_2$–$\mathcal{R}_5$ are learnable depthwise convolutions initialized with Gaussian kernels of standard deviation η_s and width k^{out}_s. Their weights are optimized through the registration objective to adjust current-scale structure features, whereas the fixed prefilters in Eq. (4) control scale transitions. At S1–S2, channel concatenation and projection yield fused matching evidence:

$$E^s = \mathcal{P}_s([E^s_{\mathrm{str}}, E^s_{\mathrm{det}}]), \qquad s \in \{1, 2\}, \tag{6}$$

where $\mathcal{P}_s$ comprises pointwise convolution, instance normalization, and nonlinear activation. At S3–S5, structure features are propagated recursively and used directly as matching evidence:

$$\begin{aligned} E^s_{\mathrm{str}} &= \mathcal{R}_s\Big(\mathcal{C}^{\mathrm{str}}_s\Big(\mathcal{A}_{\sigma^{\mathrm{str}}_{s-1}, k^{\mathrm{str}}_{s-1}}(E^{s-1}_{\mathrm{str}})\Big)\Big), \\ E^s &= E^s_{\mathrm{str}}, \qquad s \in \{3, 4, 5\}, \end{aligned} \tag{7}$$

where $\mathcal{C}^{\mathrm{str}}_s$ denotes the structure convolutional block at scale s. The operator $\mathcal{A}$ follows Eq. (4), using structure-stream parameters for the transition from scale $s-1$ to s.

3.3 Baseline-Protected Bidirectional Matcher

BPBM estimates displacement increments at S5–S2 from fixed evidence E^s_f and moving evidence $\widetilde{E}^s_m$ (Fig. 2). A one-way estimate serves as the baseline, while matching entropy and forward–backward cycle consistency control residual corrections from bidirectional candidates.

Local correspondence distributions. At each location x, we define the candidate offset set $\mathcal{N} = \{-\ell, \ldots, \ell\}^3$, where ℓ is a positive integer search radius in current-scale voxel units. Each $q \in \mathcal{N}$ specifies a three-dimensional offset, and $x + q$ is the corresponding candidate sampling location. The number of candidates is $K = |\mathcal{N}| = (2\ell + 1)^3$. The one-way branch computes the correspondence distribution p^s_{base} using dot products without channel normalization and estimates the baseline displacement $\varphi^s_{\mathrm{base}}$ as the expected candidate offset:

$$\begin{aligned} p^s_{\mathrm{base}}(q \mid x) &= \mathrm{softmax}_{q \in \mathcal{N}}\Big(\Big\langle E^s_f(x), \widetilde{E}^s_m(x+q)\Big\rangle\Big), \\ \varphi^s_{\mathrm{base}}(x) &= \sum_{q \in \mathcal{N}} p^s_{\mathrm{base}}(q \mid x)\, q, \end{aligned} \tag{8}$$

where softmax normalizes over the local candidates. The bidirectional branch uses channel-normalized features to construct fixed-to-moving and moving-to-fixed distributions:

$$\begin{aligned} p^s_{\rightarrow}(q \mid x) &= \mathrm{softmax}_{q \in \mathcal{N}}\left(\frac{\cos(E^s_f(x), \widetilde{E}^s_m(x+q))}{T}\right), \\ p^s_{\leftarrow}(q \mid x) &= \mathrm{softmax}_{q \in \mathcal{N}}\left(\frac{\cos(\widetilde{E}^s_m(x), E^s_f(x+q))}{T}\right), \end{aligned} \tag{9}$$

where $\cos$ denotes cosine similarity between channel vectors and temperature $T > 0$ controls distribution sharpness. The directional displacements are $\varphi^s_d(x) = \sum_{q \in \mathcal{N}} p^s_d(q \mid x) q$, with $d \in \{\rightarrow, \leftarrow\}$.

Bidirectional candidates and reliability. We sample the backward field at the moving locations indicated by $\varphi^s_{\rightarrow}$ to align the directional estimates and form a bidirectional candidate:

$$\widetilde{\varphi}^s_{\leftarrow} = \mathcal{W}(\varphi^s_{\leftarrow}, \varphi^s_{\rightarrow}), \qquad \varphi^s_{\mathrm{bi}} = \tfrac{1}{2}\big(\varphi^s_{\rightarrow} - \widetilde{\varphi}^s_{\leftarrow}\big). \tag{10}$$

Reliability combines local matching uncertainty and directional agreement. We measure uncertainty using normalized entropy:

$$H^s_d(x) = -\frac{1}{\log K} \sum_{q \in \mathcal{N}} p^s_d(q \mid x) \log p^s_d(q \mid x), \qquad d \in \{\rightarrow, \leftarrow\}. \tag{11}$$

The aligned bidirectional mean matching entropy $\overline{H}_s$ averages the forward entropy with the backward entropy sampled at the corresponding moving location. The cycle error e_s is the mean absolute component of the round-trip displacement:

$$\begin{aligned} \overline{H}_s &= \tfrac{1}{2}\big(H^s_{\rightarrow} + \mathcal{W}(H^s_{\leftarrow}, \varphi^s_{\rightarrow})\big), \\ e_s(x) &= \tfrac{1}{3}\big\|\varphi^s_{\rightarrow}(x) + \widetilde{\varphi}^s_{\leftarrow}(x)\big\|_1. \end{aligned} \tag{12}$$

The spatial reliability weight is

$$R_s(x) = \mathrm{clip}\big(1 - \overline{H}_s(x), 0, 1\big) \exp\left(-\frac{e_s(x)}{\tau_s}\right), \tag{13}$$

where $\mathrm{clip}(\cdot, 0, 1)$ restricts its argument to $[0, 1]$ and $\tau_s > 0$ controls cycle-error tolerance in current-scale voxel units. Concentrated matching distributions and small cycle errors yield larger weights.

Protected residual correction. Let $r_s = \varphi^s_{\mathrm{bi}} - \varphi^s_{\mathrm{base}}$ denote the residual of the bidirectional candidate relative to the one-way estimate. BPBM weights only this residual to produce the stage displacement increment:

$$\Delta\varphi_s = \varphi^s_{\mathrm{base}} + \lambda_s \, \mathrm{sg}(R_s) \odot r_s, \qquad \lambda_s = \mathrm{sigmoid}(a_s), \tag{14}$$

where a_s is an independently learned scalar at each scale and $\odot$ denotes element-wise multiplication, broadcasting spatial weights across the three displacement components. The stop-gradient operator sg acts only on R_s, preserving its forward value. Gradients remain enabled for $\varphi^s_{\mathrm{base}}$, φ^s_{bi}, and λ_s. Since $R_s(x) \in [0, 1]$ and $0 < \lambda_s < 1$, the correction satisfies $\|\Delta\varphi_s(x) - \varphi^s_{\mathrm{base}}(x)\|_2 \leq \lambda_s \|r_s(x)\|_2$. Low reliability weakens this bounded correction without directly attenuating the one-way baseline, which is recovered when $R_s(x) = 0$. At finer scales, $\Delta\varphi_s$ is composed with the upsampled accumulated displacement according to Eq. (2).

3.4 Training Objective

We train EviGuardMorph end-to-end with the objective (Balakrishnan et al., 2019):

$$\mathcal{L} = \mathcal{L}_{\mathrm{sim}}(I_f, \widehat{I}_m) + \beta \, \mathcal{L}_{\mathrm{smooth}}(\varphi_1), \tag{15}$$

where $\beta > 0$ balances the two terms. The similarity loss is $\mathcal{L}_{\mathrm{sim}}(I_f, \widehat{I}_m) = -\,\mathrm{NCC}(I_f, \widehat{I}_m)$, where NCC denotes local normalized cross-correlation. The smoothness loss penalizes squared first-order spatial finite differences of the final accumulated displacement φ_1. Training requires only image pairs, without ground-truth deformations, segmentation labels, or additional reliability supervision.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate EviGuardMorph on four public datasets covering brain MRI (**IXI** and **LPBA40** (Shattuck et al., 2008)), multi-organ abdominal CT (**AbdomenCT** (Hering et al., 2022)), and cardiac MRI (**ACDC** (Bernard et al., 2018)). LPBA40 provides a limited-data setting, while AbdomenCT includes cases with large-magnitude anatomical deformations. The first three datasets involve inter-subject registration, whereas ACDC registers end-systolic to end-diastolic images within each subject. We evaluate all methods on the same fixed test pairs. Appendix A.1 details dataset splits, preprocessing, input sizes, and pairing protocols.

Table 1: Quantitative comparison on the IXI and LPBA datasets. The best results are highlighted in **bold** and the second-best results are underlined.

Model	IXI				LPBA			
	Dice ↑	HD95 ↓	ASSD ↓	$\lvert J_\phi\rvert \leq 0$ ↓	Dice ↑	HD95 ↓	ASSD ↓	$\lvert J_\phi\rvert \leq 0$ ↓
SyN (Avants et al., 2008)	0.700 ± 0.107	3.193 ± 1.445	1.222 ± 0.576	0.000 ± 0.001	0.685 ± 0.027	6.137 ± 0.612	1.866 ± 0.181	$< \mathbf{0.0001}$
VoxelMorph-1 (Balakrishnan et al., 2019)	0.751 ± 0.033	2.759 ± 0.520	1.009 ± 0.162	0.380 ± 0.250	0.588 ± 0.026	7.217 ± 0.722	2.438 ± 0.217	0.926 ± 0.371
VoxelMorph-2 (Balakrishnan et al., 2019)	0.747 ± 0.032	2.724 ± 0.517	0.997 ± 0.157	0.367 ± 0.243	0.596 ± 0.031	7.123 ± 0.768	2.387 ± 0.245	0.587 ± 0.323
CycleMorph (Kim et al., 2021)	0.732 ± 0.037	2.894 ± 0.551	1.063 ± 0.175	0.281 ± 0.225	0.611 ± 0.029	6.966 ± 0.695	2.273 ± 0.211	0.623 ± 0.208
Im2Grid (Liu et al., 2022)	0.747 ± 0.028	2.738 ± 0.419	0.991 ± 0.132	0.043 ± 0.011	0.678 ± 0.013	6.068 ± 0.453	1.847 ± 0.097	0.015 ± 0.005
TransMorph (Chen et al., 2022)	0.758 ± 0.032	2.645 ± 0.480	0.969 ± 0.147	0.793 ± 0.277	0.700 ± 0.010	5.709 ± 0.507	1.786 ± 0.098	0.380 ± 0.244
TransMorph-Bspl (Chen et al., 2022)	0.764 ± 0.028	2.503 ± 0.415	0.938 ± 0.131	$< \mathbf{0.0001}$	0.682 ± 0.017	6.255 ± 0.574	1.910 ± 0.135	$< \mathbf{0.0001}$
TransMorph-Bayes (Chen et al., 2022)	0.759 ± 0.035	2.687 ± 0.501	0.966 ± 0.156	1.151 ± 0.304	0.659 ± 0.024	6.625 ± 0.629	2.027 ± 0.171	0.712 ± 0.190
LKU (Jia et al., 2022)	0.730 ± 0.031	2.744 ± 0.429	1.049 ± 0.149	$< \mathbf{0.0001}$	0.675 ± 0.016	5.995 ± 0.467	1.906 ± 0.115	$< \mathbf{0.0001}$
ModeT (Wang et al., 2023)	0.778 ± 0.023	2.404 ± 0.380	0.868 ± 0.113	0.014 ± 0.006	0.706 ± 0.012	5.732 ± 0.415	1.690 ± 0.095	0.006 ± 0.002
Fourier-Net (Jia et al., 2023)	0.773 ± 0.023	2.387 ± 0.424	0.886 ± 0.121	0.268 ± 0.211	0.654 ± 0.018	6.410 ± 0.530	2.023 ± 0.134	0.201 ± 0.116
WiNet (Cheng et al., 2024)	0.724 ± 0.132	3.745 ± 0.610	0.874 ± 0.116	$< \mathbf{0.0001}$	0.677 ± 0.018	6.245 ± 0.534	1.930 ± 0.134	0.020 ± 0.014
GroupMorph (Tan et al., 2024)	0.776 ± 0.022	3.474 ± 0.601	0.874 ± 0.110	0.084 ± 0.027	0.717 ± 0.009	5.507 ± 0.378	1.587 ± 0.080	0.022 ± 0.005
RDP (Wang et al., 2024)	0.777 ± 0.025	3.745 ± 0.610	0.858 ± 0.116	0.002 ± 0.003	0.723 ± 0.016	5.669 ± 0.439	1.621 ± 0.110	$< \mathbf{0.0001}$
SACB-Net (Cheng et al., 2025)	0.779 ± 0.025	3.402 ± 0.617	$\mathbf{0.841 \pm 0.121}$	0.218 ± 0.079	0.719 ± 0.011	5.559 ± 0.376	1.608 ± 0.081	0.068 ± 0.008
LessNet (Jia et al., 2025)	0.762 ± 0.029	2.566 ± 0.479	0.940 ± 0.142	0.367 ± 0.233	0.619 ± 0.022	6.839 ± 0.598	2.203 ± 0.160	0.691 ± 0.282
EviGuardMorph	$\mathbf{0.790 \pm 0.026}$	$\mathbf{2.380 \pm 0.444}$	0.856 ± 0.124	0.364 ± 0.094	$\mathbf{0.727 \pm 0.010}$	$\mathbf{5.327 \pm 0.353}$	$\mathbf{1.542 \pm 0.076}$	0.081 ± 0.011

Table 2: Quantitative comparison on the ACDC and AbdomenCT datasets. The best results are highlighted in **bold** and the second-best results are underlined.

Model	ACDC				AbdomenCT			
	Dice ↑	HD95 ↓	ASSD ↓	$\lvert J_\phi\rvert \leq 0$ ↓	Dice ↑	HD95 ↓	ASSD ↓	$\lvert J_\phi\rvert \leq 0$ ↓
SyN (Avants et al., 2008)	0.757 ± 0.122	3.551 ± 1.423	1.124 ± 0.592	$< \mathbf{0.0001}$	0.365 ± 0.039	12.467 ± 1.796	4.579 ± 0.686	0.058 ± 0.177
VoxelMorph-1 (Balakrishnan et al., 2019)	0.789 ± 0.053	3.459 ± 1.271	1.013 ± 0.286	0.040 ± 0.048	0.352 ± 0.035	12.859 ± 1.654	4.811 ± 0.647	5.254 ± 2.196
VoxelMorph-2 (Balakrishnan et al., 2019)	0.792 ± 0.053	3.526 ± 1.314	1.011 ± 0.291	0.043 ± 0.054	0.358 ± 0.032	12.793 ± 1.663	4.767 ± 0.653	4.511 ± 1.824
CycleMorph (Kim et al., 2021)	0.785 ± 0.047	3.483 ± 1.179	1.027 ± 0.259	0.022 ± 0.028	0.326 ± 0.027	13.009 ± 1.622	5.000 ± 0.614	4.391 ± 1.604
Im2Grid (Liu et al., 2022)	0.752 ± 0.059	3.641 ± 1.134	1.166 ± 0.310	$< \mathbf{0.0001}$	0.313 ± 0.035	13.552 ± 1.875	5.322 ± 0.727	0.156 ± 0.136
TransMorph (Chen et al., 2022)	0.784 ± 0.057	3.538 ± 1.333	1.055 ± 0.318	0.004 ± 0.012	0.455 ± 0.040	11.688 ± 1.893	3.869 ± 0.652	5.115 ± 1.385
TransMorph-Bspl (Chen et al., 2022)	0.789 ± 0.065	3.399 ± 1.348	0.967 ± 0.266	$< \mathbf{0.0001}$	0.454 ± 0.042	11.463 ± 1.948	3.828 ± 0.733	$< \mathbf{0.0001}$
TransMorph-Bayes (Chen et al., 2022)	0.791 ± 0.064	3.530 ± 1.341	1.022 ± 0.346	0.053 ± 0.042	0.440 ± 0.045	12.094 ± 1.839	4.072 ± 0.733	7.816 ± 1.749
LKU (Jia et al., 2022)	0.753 ± 0.059	3.688 ± 1.116	1.180 ± 0.321	$< \mathbf{0.0001}$	0.329 ± 0.044	12.831 ± 1.823	4.935 ± 0.772	$< \mathbf{0.0001}$
ModeT (Wang et al., 2023)	0.796 ± 0.044	3.360 ± 1.159	0.966 ± 0.254	0.001 ± 0.001	0.523 ± 0.038	10.287 ± 1.608	3.109 ± 0.469	0.561 ± 0.289
Fourier-Net (Jia et al., 2023)	0.776 ± 0.048	3.416 ± 1.116	1.039 ± 0.252	0.007 ± 0.020	0.377 ± 0.040	12.322 ± 1.633	4.475 ± 0.659	6.278 ± 3.416
WiNet (Cheng et al., 2024)	0.780 ± 0.048	3.460 ± 1.173	1.060 ± 0.269	$< \mathbf{0.0001}$	0.403 ± 0.042	12.238 ± 2.014	4.356 ± 0.766	$< \mathbf{0.0001}$
GroupMorph (Tan et al., 2024)	0.797 ± 0.050	3.434 ± 1.210	0.971 ± 0.282	0.014 ± 0.017	0.554 ± 0.040	9.833 ± 1.739	2.866 ± 0.447	0.915 ± 0.449
RDP (Wang et al., 2024)	0.799 ± 0.046	3.367 ± 1.223	0.938 ± 0.253	$< \mathbf{0.0001}$	0.557 ± 0.056	9.869 ± 1.835	2.858 ± 0.581	0.337 ± 0.219
SACB-Net (Cheng et al., 2025)	0.798 ± 0.048	3.335 ± 1.199	0.933 ± 0.259	0.070 ± 0.042	0.571 ± 0.050	9.711 ± 1.795	2.700 ± 0.480	3.071 ± 0.776
LessNet (Jia et al., 2025)	0.781 ± 0.054	3.542 ± 1.238	1.039 ± 0.283	$< \mathbf{0.0001}$	0.340 ± 0.031	12.914 ± 1.637	4.889 ± 0.600	$< \mathbf{0.0001}$
EviGuardMorph (Ours)	$\mathbf{0.812 \pm 0.041}$	$\mathbf{3.261 \pm 1.201}$	$\mathbf{0.902 \pm 0.243}$	0.016 ± 0.018	$\mathbf{0.591 \pm 0.034}$	$\mathbf{9.302 \pm 1.784}$	$\mathbf{2.586 \pm 0.471}$	2.411 ± 0.624

Implementation details. We implement EviGuardMorph in PyTorch and train it on an NVIDIA RTX 4090 GPU using Adam (Kingma & Ba, 2014) for 500 epochs with a batch size of 1. The initial learning rate is 10^{-4}, with polynomial decay of power 0.9. Weight decay and the smoothness weight are 10^{-4} and $\beta = 0.1$, respectively. BPBM uses a $3 \times 3 \times 3$ matching window and temperature $T = 0.1$. Detailed scale-wise filtering and matching configurations are provided in Appendix A.1.

4.2 Comparison with Existing Methods

We compare EviGuardMorph with optimization-based and learning-based methods (Tables 1 and 2). It achieves the highest mean Dice on IXI, LPBA, ACDC, and AbdomenCT (0.790, 0.727, 0.812, and 0.591), exceeding the second-best results by approximately 1.1, 0.4, 1.3, and 2.0 percentage points, respectively. The largest gain on AbdomenCT supports its effectiveness under large inter-subject deformations, while the LPBA improvement demonstrates effectiveness with only 25 training pairs. EviGuardMorph also achieves the lowest mean HD95 on all four datasets and the lowest mean ASSD on LPBA, ACDC, and AbdomenCT, with the second-lowest mean ASSD on IXI. These results further support accurate anatomical boundary alignment. Paired Wilcoxon signed-rank tests on Dice against the strongest competing method per dataset confirm significant differences (Table 3), with $p < 0.001$ on IXI, ACDC, and AbdomenCT, and $p = 0.0359$ on LPBA. Appendix A.3 (Figs. 7–10) provides complementary registration and frequency-response visualizations. These visualizations show that EviGuardMorph achieves close alignment of anatomical contours and lower out-of-band-to-in-band RMS response ratios at the S3-to-S4 transition, along with spatially structured in-band responses and detailed S2 features. These observations support the FBE objective of suppressing coarse-scale out-of-band interference while preserving fine-scale localization cues. These observations support the FBE objective of suppressing coarse-scale out-of-band interference while preserving fine-scale localization cues. Appendix A.5 further examines local anatomical mismatches that remain after registration.

Table 3: Wilcoxon signed-rank tests on Dice against the strongest baseline per dataset.

Dataset	Method	p-value	Sig.
IXI	SACB-Net	1.14×10^{-10}	$***$
LPBA	RDP	3.59×10^{-2}	$*$
ACDC	RDP	8.01×10^{-4}	$***$
AbdomenCT	SACB-Net	1.46×10^{-5}	$***$

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table 4: Ablation study of the proposed FBE and BPBM modules on the LPBA and AbdomenCT datasets. The best results are highlighted in bold and the second-best results are underlined.

FBE	BPBM	LPBA		AbdomenCT	
		Dice ↑	$\lvert J_\phi\rvert \le 0$ ↓	Dice ↑	$\lvert J_\phi\rvert \le 0$ ↓
×	×	0.7039 ± 0.0122	0.1063 ± 0.0178	0.5467 ± 0.0378	4.4382 ± 0.8145
✓	×	$\underline{0.7267 \pm 0.0109}$	0.0882 ± 0.0145	$\underline{0.5839 \pm 0.0360}$	$\mathbf{2.1077 \pm 0.6339}$
×	✓	0.7235 ± 0.0099	$\mathbf{0.0628 \pm 0.0116}$	0.5817 ± 0.0385	$\underline{2.3396 \pm 0.6220}$
✓	✓	$\mathbf{0.7275 \pm 0.0105}$	$\underline{0.0806 \pm 0.0109}$	$\mathbf{0.5905 \pm 0.0345}$	2.4115 ± 0.6244

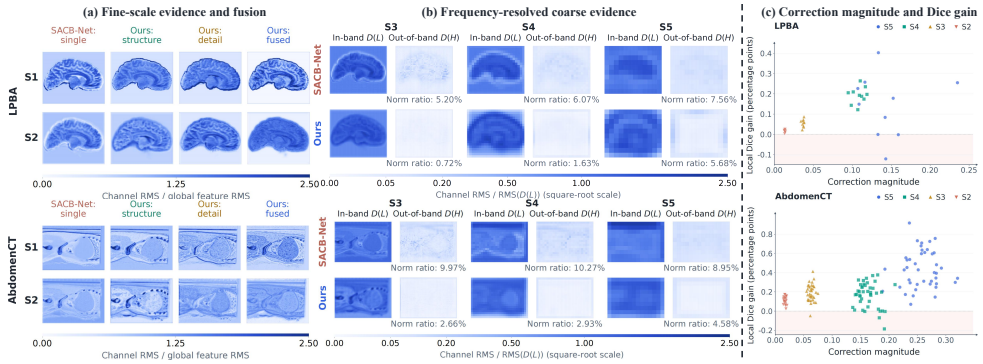


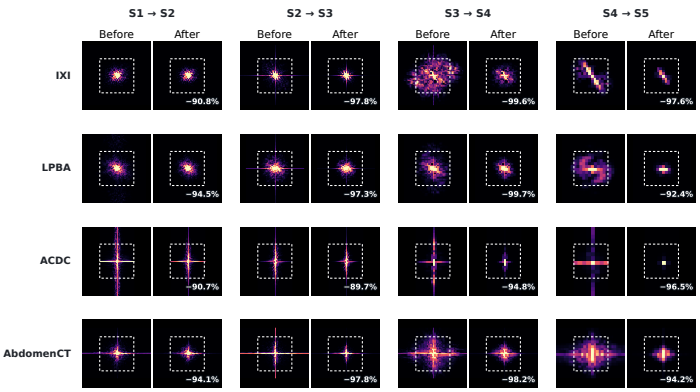
Figure 3: **Evidence visualization and scale-wise correction analysis.** Rows show LPBA (top) and AbdomenCT (bottom). (a) S1–S2 evidence from SACB-Net and FBE's structure, detail, and fused streams, visualized as channel RMS normalized by global feature RMS. (b) In-band $D(L)$ and out-of-band $D(H)$ downsampling responses at S3–S5, normalized by $\mathrm{RMS}(D(L))$ and displayed with a shared square-root color scale. Percentages indicate the volumetric norm ratio $\|D(H)\|_2/\|D(L)\|_2$. (c) Correction magnitude (input voxels) versus local Dice gain (percentage points) at S2–S5. Each point represents one registration pair at one scale.

4.3 Ablation Study

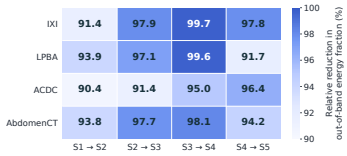
Component contributions and visual analysis. Table 4 isolates the contributions of FBE and BPBM on LPBA and AbdomenCT. Relative to the configuration without either module, FBE alone improves mean Dice by 2.28 and 3.72 percentage points, respectively. At S1–S2, the structure stream captures anatomical contours, while the detail stream retains boundary and texture variations. Their fusion preserves both (Fig. 3(a)). At S3, the out-of-band-to-in-band response norm ratio decreases from 5.20% to 0.72% on LPBA and from 9.97% to 2.66% on AbdomenCT compared with SACB-Net (Fig. 3(b)). These observations support scale-adapted evidence construction that preserves fine-scale localization cues while limiting coarse-scale out-of-band interference.

BPBM alone improves mean Dice by 1.96 and 3.50 percentage points on LPBA and AbdomenCT, respectively. S5 generally yields larger local Dice gains than S2, particularly on AbdomenCT (Fig. 3(c)). Coarse corrections address broader structural misalignment and guide subsequent pre-alignment, whereas S2 refines residual local discrepancies. Early errors can propagate, making reliable coarse-scale correction important. The variable S5 gains on LPBA motivate reliability-weighted corrections that preserve the one-way baseline. Combining FBE and BPBM achieves the highest mean Dice across ablations on both datasets, reaching 0.7275 on LPBA and 0.5905 on AbdomenCT. The full model also reduces folding relative to the configuration without either module. These support their complementary roles in evidence construction and displacement estimation.

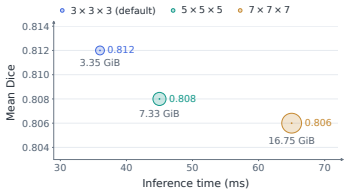
Spectral analysis of prefiltering. We further examine changes in out-of-band components before scale transitions. Spectra from one sample per dataset show that prefiltering attenuates components outside the new Nyquist passband (Fig. 4(a)). Aggregated results across four datasets and four scale transitions show a 90.4%–99.7% relative reduction in the out-of-band spectral energy fraction before downsampling (Fig. 4(b)). Consistent with the reduced coarse-scale out-of-band responses in Fig. 3(b), prefiltering decreases the relative contribution of out-of-band components before the sampling rate is reduced, helping mitigate the risk of aliasing during subsequent downsampling. Controlled ablations on LPBA and ACDC further show that pre-downsampling filtering within FBE yields higher mean Dice than removing fixed prefiltering or filtering after pooling (Appendix A.2).



(a) Spectral attenuation



(b) Out-of-band suppression



(c) Matching-window sensitivity

Figure 4: **Spectral analysis and matching-window sensitivity.** (a) Spectra before and after Gaussian prefiltering across pyramid transitions, using one sample per dataset. Dashed boxes mark the post-downsampling Nyquist passband. (b) Relative reduction in out-of-band spectral energy fraction across datasets and transitions. (c) ACDC mean Dice versus inference time. Colors encode window size; bubble area indicates peak training GPU memory (GiB).

Table 5: Sensitivity to scale-wise λ initialization on LPBA. Best results are **bold**.

Strategy	Initial λ (S5/S4/S3/S2)	Final effective λ (S5/S4/S3/S2)	Dice ↑	$\lvert J_\phi\rvert \le 0$ ↓
Uniform	0.0500/0.0500/0.0500/0.0500	0.0877/0.0879/0.0880/0.0874	0.7245 ± 0.0110	0.0936 ± 0.0141
Default	0.0500/0.1000/0.1000/0.1000	0.1009/0.1909/0.1916/0.1885	$\mathbf{0.7275 \pm 0.0105}$	$\mathbf{0.0806 \pm 0.0109}$
Middle-focused	0.0010/0.1000/0.0500/0.0010	0.0022/0.1919/0.1019/0.0022	0.7252 ± 0.0097	0.0849 ± 0.0152

Table 6: Sensitivity to scale-wise τ schedules on LPBA. Best results are **bold**.

Strategy	τ (S5/S4/S3/S2)	Dice ↑	$\lvert J_\phi\rvert \le 0$ ↓
Reverse	0.5000/1.0000/1.5000/2.0000	0.7255 ± 0.0114	0.0823 ± 0.0112
Uniform	1.0000/1.0000/1.0000/1.0000	0.7260 ± 0.0110	$\mathbf{0.0784 \pm 0.0126}$
Default	2.0000/1.5000/1.0000/0.5000	$\mathbf{0.7275 \pm 0.0105}$	0.0806 ± 0.0109

Parameter sensitivity. On LPBA, the default correction-strength initialization, smaller at S5 than at S4–S2, achieves the highest mean Dice (0.7275) and lowest folding percentage (0.0806%) among the tested strategies (Table 5). Distinct learned strengths indicate that training retains sensitivity to initialization. For cycle error tolerance, decreasing the tolerance from S5 to S2 improves Dice over reverse and uniform schedules, although uniform tolerance yields slightly less folding (0.0784% versus 0.0806%) (Table 6). These results support the default initialization and more permissive coarse-scale, but stricter fine-scale, reliability weighting of cycle-inconsistent candidates.

On ACDC, the default $3 \times 3 \times 3$ window achieves the highest mean Dice (0.812), with 36 ms inference time and 3.35 GiB peak training GPU memory (Fig. 4(c)). Windows of $5 \times 5 \times 5$ and $7 \times 7 \times 7$ reduce Dice to 0.808 and 0.806, while increasing inference time to 45 and 65 ms and peak training GPU memory to 7.33 and 16.75 GiB, respectively.

5 Conclusion

We presented EviGuardMorph, a deformable medical image registration framework that combines scale-adaptive evidence construction with reliability-guided displacement updates. FBE suppresses coarse-scale out-of-band interference while retaining fine-scale localization cues. Spectral analysis shows that prefiltering yields a 90.4%–99.7% relative reduction in the out-of-band spectral energy fraction before downsampling. BPBM incorporates bidirectional evidence through bounded, reliability-weighted residual corrections without attenuating the one-way estimate. Experiments on four datasets demonstrate leading registration accuracy, while ablations support the complementary roles of evidence construction and controlled use in anatomical alignment.

AI USE STATEMENT

In this work, we used generative AI tools solely for grammatical editing of the manuscript. All edits were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

The architecture and training objective of EviGuardMorph are described in Sections 3.1–3.4. Appendix A.1 provides the dataset splits, pairing protocols, network and training configurations, evaluation metrics, and checkpoint selection criteria. Appendix A.3 details the frequency-response and scale-wise correction analyses. We will publicly release the source code upon acceptance of the paper.

REFERENCES

Mazdak Abulnaga, Andrew Hoopes, Malte Hoffmann, Robin Magnet, Maks Ovsjanikov, Lilla Zollei, John Guttag, Bruce Fischl, and Adrian Dalca. Unified brain surface and volume registration. In *International Conference on Learning Representations*, volume 2026, pp. 132177–132193, 2026.

Brian B. Avants, Charles L. Epstein, Murray Grossman, and James C. Gee. Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 2008.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: a learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525, 2018.

Junyu Chen, Eric C. Frey, Yufan He, William P. Segars, Ye Li, and Yong Du. TransMorph: transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022.

Linwei Chen, Lin Gu, and Ying Fu. When semantic segmentation meets frequency aliasing. In *International Conference on Learning Representations*, volume 2024, pp. 43194–43216, 2024.

Xinxing Cheng, Xi Jia, Wenqi Lu, Qiufu Li, Linlin Shen, Alexander Krull, and Jinming Duan. WiNet: wavelet-based incremental learning for efficient medical image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 761–771, 2024.

Xinxing Cheng, Tianyang Zhang, Wenqi Lu, Qingjie Meng, Alejandro F. Frangi, and Jinming Duan. SACB-Net: spatial-awareness convolutions for medical image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5227–5237, 2025.

Xueming Fu, Fenghe Tang, Rongsheng Wang, Yingtai Li, Lixia Han, Jian Lu, Zihang Jiang, and S. Kevin Zhou. MedGMAE: Gaussian masked autoencoders for medical volumetric representation learning. In *International Conference on Learning Representations*, 2026.

Qi Han, Yuxuan Cai, and Xiangyu Zhang. RevColV2: exploring disentangled representations in masked image modeling. *Advances in Neural Information Processing Systems*, 36:29273–29291, 2023.

Alessa Hering, Lasse Hansen, Tony C. W. Mok, Albert C. S. Chung, Hanna Siebert, Stephanie Häger, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, et al. Learn2Reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2022.

Rohit Jena, Vedant Zope, Pratik A. Chaudhari, and James Gee. A scalable distributed framework for multimodal GigaVoxel image registration. In *International Conference on Learning Representations*, volume 2026, pp. 119573–119592, 2026.

Xi Jia, Joseph Bartlett, Tianyang Zhang, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. U-Net vs transformer: is U-Net outdated in medical image registration? In *International Workshop on Machine Learning in Medical Imaging*, pp. 151–160, 2022.

Xi Jia, Joseph Bartlett, Wei Chen, Siyang Song, Tianyang Zhang, Xinxing Cheng, Wenqi Lu, Zhaowen Qiu, and Jinming Duan. Fourier-Net: fast image registration with band-limited deformation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1015–1023, 2023.

Xi Jia, Wenqi Lu, Xinxing Cheng, and Jinming Duan. Decoder-only image registration. *IEEE Transactions on Medical Imaging*, 44(8):3356–3369, 2025.

Boah Kim, Dong Hwan Kim, Seong Ho Park, Jieun Kim, June-Goo Lee, and Jong Chul Ye. CycleMorph: cycle consistent unsupervised deformable image registration. *Medical Image Analysis*, 71:102036, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Yihao Liu, Lianrui Zuo, Shuo Han, Yuan Xue, Jerry L. Prince, and Aaron Carass. Coordinate translator for learning deformable medical image registration. In *International Workshop on Multiscale Multimodal Medical Imaging*, pp. 98–109, 2022.

Yu Bing Luo, Nian Shi, Jia Qin, Zekai Ji, Pinle Qin, Jianchao Zeng, and Jianghui Cai. Revisiting downsampling in semantic segmentation: fighting aliasing with dynamic Gaussian and Gabor frequency filters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7707–7715, 2026.

Md Ashiqur Rahman and Raymond A. Yeh. Group downsampling with equivariant anti-aliasing. In *International Conference on Learning Representations*, volume 2025, pp. 37461–37491, 2025.

Daniel Rueckert, Luke I. Sonoda, Carmel Hayes, Derek L. G. Hill, Martin O. Leach, and David J. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

David W. Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L. Narr, Russell A. Poldrack, Robert M. Bilder, and Arthur W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064–1080, 2008.

Zuopeng Tan, Lihe Zhang, Yanan Lv, Yili Ma, and Huchuan Lu. GroupMorph: medical image registration via grouping network with contextual fusion. *IEEE Transactions on Medical Imaging*, 43(11):3807–3819, 2024.

Haiqiao Wang, Dong Ni, and Yi Wang. ModeT: learning deformable image registration via motion decomposition transformer. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 740–749, 2023.

Haiqiao Wang, Dong Ni, and Yi Wang. Recursive deformable pyramid network for unsupervised medical image registration. *IEEE Transactions on Medical Imaging*, 43(6):2229–2240, 2024.

A Appendix

This appendix provides implementation and evaluation details, controlled FBE ablations, qualitative and frequency-response analyses, structure-wise registration results, and residual mismatch analysis. It is organized as follows.

A.1 Implementation and Evaluation Details

1. Datasets and pairing protocols
2. Network configuration
3. Evaluation metrics
4. Model selection and reporting

A.2 Controlled Ablations of FBE

1. Experimental setup
2. Effect of filter placement
3. Filtering and evidence construction
4. Alignment accuracy and deformation regularity

A.3 Qualitative Comparisons and Frequency-Response Analysis

1. Frequency decomposition and analysis
2. Native frequency responses and visualization
3. Prefilter spectral attenuation
4. Scale-wise correction analysis
5. Qualitative comparisons

A.4 Structure-wise Registration Performance

1. Brain structures
2. Cardiac structures
3. Abdominal structures

A.5 Limitations Analysis

A.1 Implementation and Evaluation Details

Datasets and pairing protocols. The four datasets assess complementary registration settings. IXI evaluates inter-subject brain registration, while LPBA examines performance with limited training data. AbdomenCT tests registration under large inter-subject anatomical deformations, and ACDC evaluates intra-subject alignment between cardiac phases. Table 7 summarizes the dataset splits, input sizes, and registration-pair counts. All learning-based methods use the same training pairs within each dataset, and all methods are evaluated on the same fixed test pairs.

Table 7: Dataset splits, input sizes, and registration-pair counts. Case counts and pair counts are reported separately.

Dataset	Cases (train/val/test)	Input size (voxels)	Registration pairs (train/test)
IXI	403/58/115	$160 \times 192 \times 224$	403/115
LPBA	25/5/10	$160 \times 192 \times 160$	25/10
AbdomenCT	20/3/7	$192 \times 160 \times 224$	380/42
ACDC	105/15/30	$128 \times 128 \times 64$	105/30

IXI and LPBA are used for inter-subject brain registration. IXI images undergo skull stripping and affine alignment to MNI152 space, with Dice evaluated over 27 foreground labels from FreeSurfer-derived segmentations. LPBA40 contains 40 brain MRI volumes annotated with 56 anatomical structures. Dice evaluation uses 54 foreground labels, excluding the cerebellum and brainstem. Left and right structures are evaluated separately and averaged into 27 bilateral regions only for the LPBA structure-wise visualizations. For both datasets, moving–fixed pairs are randomly sampled between distinct subjects within each split. The pair lists are generated once and remain fixed throughout training and evaluation, yielding 403/115 training/test pairs for IXI and 25/10 for LPBA.

AbdomenCT is used for inter-subject abdominal registration, with Dice evaluated over 13 foreground anatomical labels. All ordered pairs of distinct scans are constructed within each split, with

opposite registration directions treated as separate pairs. The 20 training scans yield $20 \times 19 = 380$ pairs, and the seven test scans yield $7 \times 6 = 42$ pairs.

ACDC is used for intra-subject registration across cardiac phases. Each subject contributes one pair, with the end-systolic image as moving and the end-diastolic image as fixed. The training and test sets contain 105 and 30 pairs, respectively. Dice is evaluated over three foreground labels corresponding to the right ventricle, myocardium, and left ventricle.

Across all four datasets, Dice is computed separately for each evaluated foreground label and averaged with equal weight within each registration pair. Background is excluded.

Network configuration. The pyramid uses $\rho = 2$. Fixed structure-stream prefilters use $(\sigma_s^{\mathrm{str}}, k_s^{\mathrm{str}}) = (0.8, 5), (1.2, 5), (1.5, 5), (2.0, 7)$ for transitions S1–S2 through S4–S5, respectively, while the detail stream uses $(\sigma_1^{\mathrm{det}}, k_1^{\mathrm{det}}) = (0.5, 3)$ for S1–S2. Structure filtering at S2–S5 uses learnable operators $\mathcal{R}_s$ with Gaussian initializations $(\eta_s, k_s^{\mathrm{out}}) = (0.8, 3), (1.2, 5), (1.5, 5), (2.0, 7)$, respectively, with learnable weights rather than fixed kernels. BPBM uses a $3 \times 3 \times 3$ window ($\ell = 1$, $K = 27$) and $T = 0.1$. From S5 to S2, cycle-error tolerances are fixed at $\tau_s = (2.0, 1.5, 1.0, 0.5)$ in current-scale voxel units, and learnable correction strengths λ_s are initialized to $(0.05, 0.10, 0.10, 0.10)$. Gradients through reliability weights are detached. SRRH applies fixed Gaussian smoothing with $\sigma_{\mathrm{ref}} = 1.5$ and $k_{\mathrm{ref}} = 5$. All Gaussian standard deviations are expressed in voxel units of their respective grids. In the ablations, removing FBE replaces it with the convolutional encoder used in SACB-Net. Removing BPBM replaces it with SACB-Net's local attention-based matching. All four configurations retain the same SRRH, deformation composition, training schedule, and checkpoint selection criterion.

Evaluation metrics. We evaluate anatomical alignment using Dice, HD95, and average symmetric surface distance (ASSD), and deformation regularity using $|J_\phi| \leq 0$. Dice measures overlap between warped moving and fixed segmentations, averaged over the evaluated structures within each registration pair. HD95 measures the 95th-percentile Hausdorff distance between corresponding anatomical boundaries, while ASSD measures their average bidirectional surface distance. Both distances are reported in voxel units. For the final transformation $\phi = \mathrm{Id} + \varphi_1$, $J_\phi = \nabla\phi$ denotes its Jacobian matrix, with $|J_\phi|$ denoting the determinant. The metric $|J_\phi| \leq 0$ reports the percentage of evaluated voxels with non-positive determinants. Higher Dice and lower HD95 and ASSD indicate better anatomical alignment, while a lower folding percentage indicates fewer local foldings.

Model selection and reporting. For each learning-based method, the checkpoint with the highest validation Dice is evaluated on the fixed test pairs. EviGuardMorph is trained using image pairs alone; segmentation labels are used only for validation and testing. Metrics are computed for each registration pair and summarized as the mean and standard deviation across test pairs. In the matching-window analysis, runtime denotes inference time, whereas GPU memory denotes peak usage during training.

A.2 Controlled Ablations of FBE

Experimental setup. To evaluate the contribution of pre-downsampling anti-aliasing in FBE, we conduct controlled filtering ablations on LPBA and ACDC. All configurations retain BPBM. Within the full FBE architecture, we compare the removal of fixed Gaussian prefiltering, filtering before downsampling, and filtering after downsampling while keeping the remaining FBE components unchanged. We also compare single-stream encoders with and without prefiltering to examine its relationship with evidence construction. Table 8 reports the mean Dice and folding rates, and Fig. 5 shows the changes in Dice relative to the baseline without FBE.

Effect of filter placement. Within the full FBE architecture, pre-downsampling filtering achieves the highest mean Dice on both datasets, reaching 0.7275 on LPBA and 0.8120 on ACDC. Removing fixed prefiltering reduces these values to 0.7265 and 0.8110, while moving filtering after pooling yields 0.7261 and 0.8116, respectively. Pre-downsampling filtering therefore improves mean Dice by 0.10 percentage points on both datasets over removing fixed prefiltering, and by 0.14 and 0.04 percentage points over post-pooling filtering. This pattern is consistent with controlling out-of-band

Table 8: Controlled encoder and filtering ablations on LPBA and ACDC. All variants retain BPBM. Best results are in **bold**.

Configuration	LPBA		ACDC	
	DSC ↑	Folding (%) ↓	DSC ↑	Folding (%) ↓
w/o FBE	0.7235 ± 0.0099	$\mathbf{0.0628 \pm 0.0116}$	0.8035 ± 0.0418	0.0210 ± 0.0185
w/o FBE, w/ prefiltering	0.7208 ± 0.0115	0.0764 ± 0.0133	0.8070 ± 0.0428	$\mathbf{0.0029 \pm 0.0041}$
w/o fixed prefiltering	0.7265 ± 0.0108	0.0752 ± 0.0134	0.8110 ± 0.0404	0.0129 ± 0.0165
w/ fixed post-pooling filtering	0.7261 ± 0.0114	0.0784 ± 0.0113	0.8116 ± 0.0408	0.0154 ± 0.0189
Full model	$\mathbf{0.7275 \pm 0.0105}$	0.0806 ± 0.0109	$\mathbf{0.8120 \pm 0.0410}$	0.0160 ± 0.0180

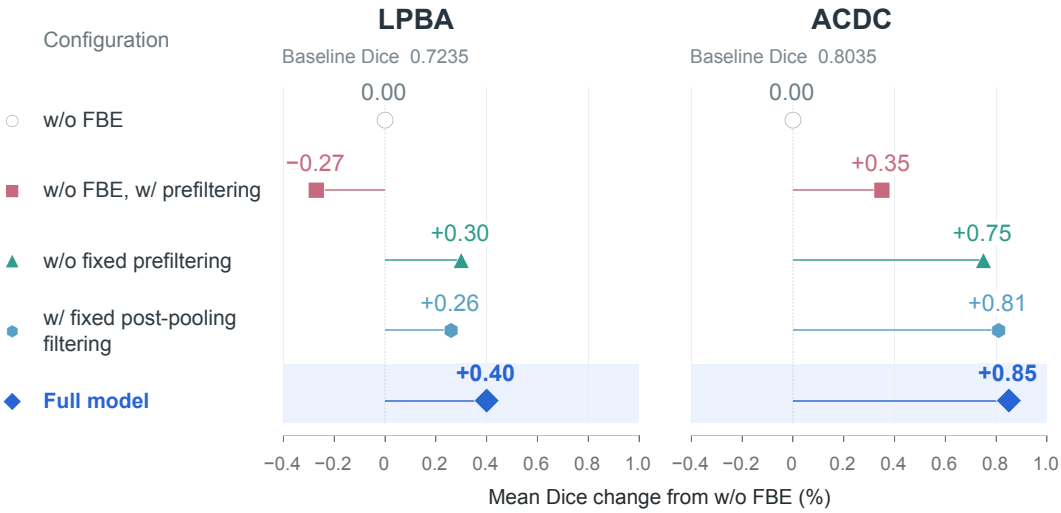


Figure 5: **Controlled ablations of FBE and fixed Gaussian filtering.** Mean Dice changes on LPBA and ACDC are measured relative to the single-stream baseline without FBE. BPBM is retained in all configurations. Values are reported in percentage points, with horizontal segments connecting each mean difference to zero. The full model applies fixed Gaussian prefiltering before pooling.

components before reducing the sampling rate and supports the chosen filter placement within the FBE.

Filtering and evidence construction. The single-stream results show that adding prefiltering alone does not consistently improve performance across the two tasks. It reduces mean Dice on LPBA from 0.7235 to 0.7208 but improves ACDC from 0.8035 to 0.8070. This difference may reflect task-dependent requirements for detail preservation and smoothing. Smoothing single-stream features may weaken local boundary cues for fine-grained brain registration on LPBA, whereas cardiac inter-phase registration on ACDC may benefit from smoother structural evidence. In contrast, full FBE outperforms the single-stream encoder with prefiltering by 0.67 and 0.50 percentage points, respectively. These results support combining bandwidth control with joint structure and detail evidence construction, rather than treating fixed low-pass filtering as a standalone performance enhancement.

Alignment accuracy and deformation regularity. FBE without fixed prefiltering also outperforms the single-stream baseline, indicating that fixed prefiltering is not the sole source of the full model's gains. With BPBM retained, full FBE improves mean Dice over the baseline without FBE by 0.40 percentage points on LPBA and 0.85 percentage points on ACDC. These gains show that improved evidence construction remains beneficial in the presence of reliability-weighted matching, consistent with the complementary roles of FBE in constructing correspondence evidence and BPBM in constraining displacement corrections. Alignment accuracy also exhibits a trade-off with deformation regularity. On LPBA, the full model has a folding rate of 0.0806%, compared with 0.0628% for the baseline without FBE. On ACDC, its folding rate decreases from the baseline's 0.0210% to 0.0160%, but remains above the 0.0029% achieved by the single-stream encoder with

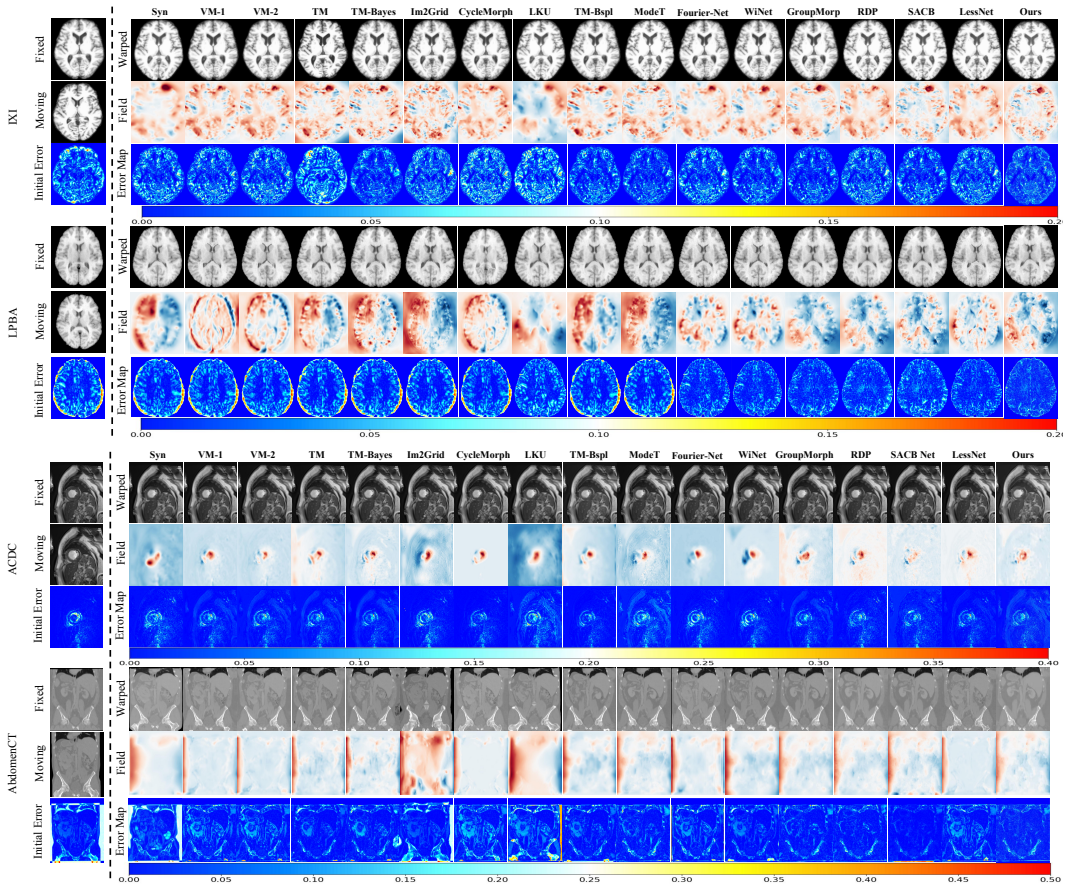


Figure 6: **Qualitative registration comparison on four datasets.** Top: IXI and LPBA. Bottom: ACDC and AbdomenCT. For each dataset, the first column shows the fixed image, moving image, and initial absolute error. Subsequent columns show the warped image, deformation field visualization, and absolute error map for each method. Lower absolute errors indicate smaller intensity discrepancies.

prefiltering. The full model therefore achieves the highest mean Dice, but not the lowest folding rates.

Taken together, the preceding spectral analysis shows that prefiltering attenuates out-of-band components before downsampling, while these controlled experiments show additional registration accuracy gains from this operation within the full FBE architecture. Both findings support the design of FBE: preserving fine-scale localization cues while controlling bandwidth before downsampling to construct structural evidence for coarse-scale matching.

A.3 Qualitative Comparisons and Frequency-Response Analysis

Fig. 6 compares registered images and intensity residuals across the four datasets. Figs. 7–10 complement these views with local segmentation contours, deformation grids, and feature responses.

Frequency decomposition and analysis. Downsampling reduces the alias-free frequency range, allowing insufficiently attenuated out-of-band components to fold into coarse-scale representations. To examine their contribution to coarse-scale evidence, we decompose the input features and evaluate their responses through each model's native downsampling operator. The six-method comparison uses the S3-to-S4 transition, reducing spatial resolution from one-quarter to one-eighth of the input image resolution along each axis. For downsampling factor $\rho = 2$, the resulting Nyquist limit is

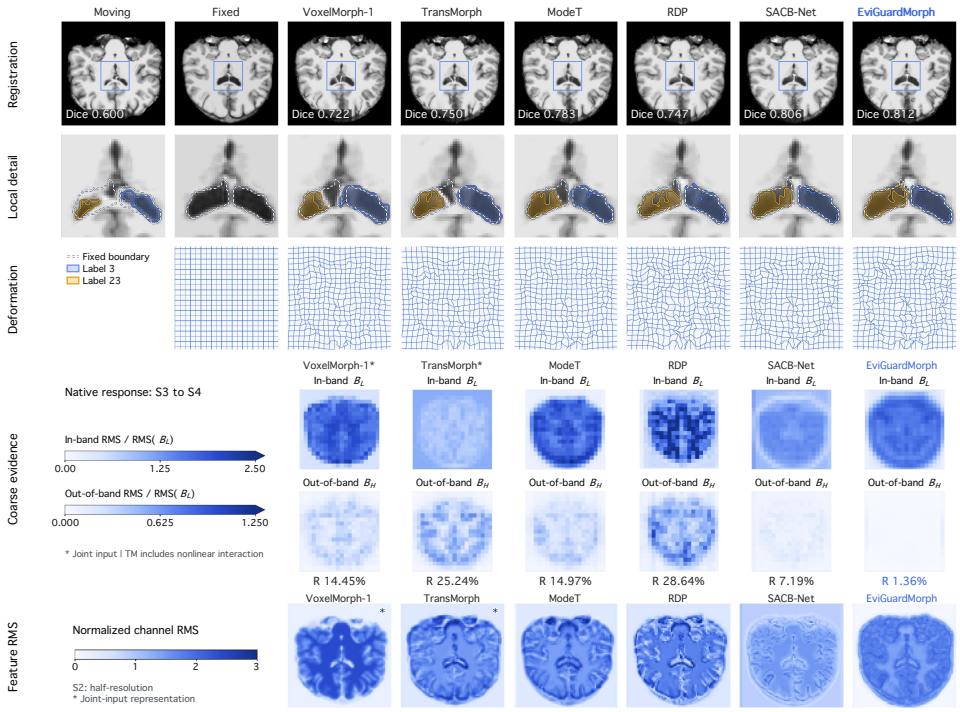


Figure 7: Example registration and feature-response comparison on IXI.

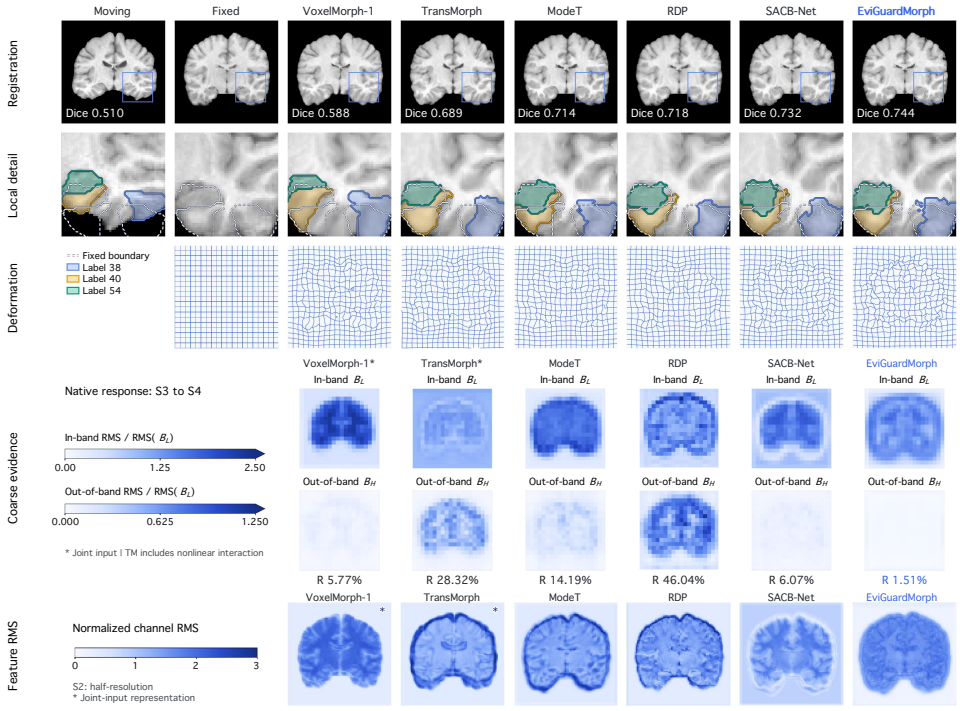


Figure 8: Example registration and feature-response comparison on LPBA.

$1/(2\rho) = 0.25$ cycles per input-feature voxel. The corresponding passband mask is

$$M_\rho(\mathbf{f}) = \mathbb{I}\left[\max_{i\in\{1,2,3\}} |f_i| \le \frac{1}{2\rho}\right], \tag{16}$$

where $\mathbf{f}$ denotes spatial frequency and $\mathbb{I}$ is the indicator function. We apply eight-voxel reflection padding on each side of the input features X and compute an orthonormal three-dimensional Fourier transform independently for each channel. The reconstructed in-band component L and out-of-band residual H are

$$L = \mathcal{C}\left\{\mathcal{F}^{-1}[M_\rho \odot \mathcal{F}(\mathcal{P}X)]\right\}, \qquad H = X - L, \tag{17}$$

where $\mathcal{P}$ and $\mathcal{C}$ denote reflection padding and cropping to the original dimensions, respectively, and $\mathcal{F}$ denotes the Fourier transform. Feature means are retained, and no additional window function is applied. Reflection padding is used for this decomposition, while the subsequent transition operator retains its native boundary handling.

Native frequency responses and visualization. Let D denote the native downsampling operator. For VoxelMorph-1 and RDP, D is the stride-two convolution before activation. For TransMorph, it is Patch Merging with LayerNorm and linear projection. ModeT and SACB-Net use average pooling, whereas EviGuardMorph uses Gaussian prefiltering followed by average pooling. VoxelMorph-1 and TransMorph operate on jointly encoded image-pair features, while the other methods are analyzed using fixed-image features. We remove the zero-input response and define

$$B_L = D(L) - D(0), \qquad B_H = D(X) - D(L). \tag{18}$$

Here, B_L is the bias-corrected in-band response, and B_H is the response increment induced by adding H to L. They satisfy $D(X) - D(0) = B_L + B_H$. For linear or affine operators, $B_H = D(H) - D(0)$. For TransMorph, nonlinear normalization also introduces interactions between L and H into B_H. We quantify the relative out-of-band contribution using the root-mean-square (RMS) response ratio

$$R = \frac{\mathrm{RMS}(B_H)}{\mathrm{RMS}(B_L)} \times 100\%. \tag{19}$$

RMS is computed over all channels and spatial locations in the three-dimensional response volume. Thus, R measures the relative response induced by out-of-band inputs, rather than the high-frequency energy fraction of the output spectrum.

The S3–S5 analyses in Figs. 1(b) and 3(b) refer to the target scales of the S2-to-S3, S3-to-S4, and S4-to-S5 transitions, respectively. For the linear, zero-bias operators shown there, $D(0) = 0$, yielding $B_L = D(L)$ and $B_H = D(H)$. Their volumetric L_2 norm ratios equal the corresponding RMS ratios and are reported as percentages.

For Figs. 7–10, we compute channel RMS at each spatial location in B_L and B_H and normalize both maps by the global RMS of B_L. Shared color limits are $[0, 2.5]$ for in-band responses and $[0, 1.25]$ for out-of-band responses across methods. Display clipping does not affect R. The $D(L)/D(H)$ maps in Fig. 3(b) use the shared square-root color scale specified in its caption. Fine-scale feature maps show channel RMS normalized by the global RMS of each corresponding feature volume, revealing spatial response patterns independently of overall feature amplitude.

Prefilter spectral attenuation. Fig. 4(a,b) evaluates spectral attenuation before spatial downsampling. For a structure-stream feature Z, let $\widetilde{Z} = G_{\sigma,k} *_{\mathrm{dw}} Z$ denote its Gaussian-prefiltered counterpart. We define the channel-summed power spectrum as $P_Z(\mathbf{f}) = \sum_c |\mathcal{F}(Z_c)(\mathbf{f})|^2$. The out-of-band energy fraction and its relative reduction are

$$\begin{aligned} \mathcal{E}_{\mathrm{out}}(Z) &= \frac{\sum_{\mathbf{f}} [1 - M_\rho(\mathbf{f})] P_Z(\mathbf{f})}{\sum_{\mathbf{f}} P_Z(\mathbf{f})}, \\ \mathrm{Reduction} &= \frac{\mathcal{E}_{\mathrm{out}}(Z) - \mathcal{E}_{\mathrm{out}}(\widetilde{Z})}{\mathcal{E}_{\mathrm{out}}(Z)} \times 100\%. \end{aligned} \tag{20}$$

Denominators are floored at 10^{-12} for numerical stability. Both measurements precede pooling and use identical frequency grids and cutoffs, summing power over all channels and spatial frequencies. Unlike the native-response decomposition, this calculation uses the original feature dimensions without reflection padding. Feature means are retained, and no window function is applied. For each dataset and pyramid transition, relative reductions are computed per image and averaged over the full test set. Panel (a) displays example spectra, whereas panel (b) reports the corresponding test-set averages. This energy-based measure characterizes prefilter attenuation and is distinct from the native response ratio R.

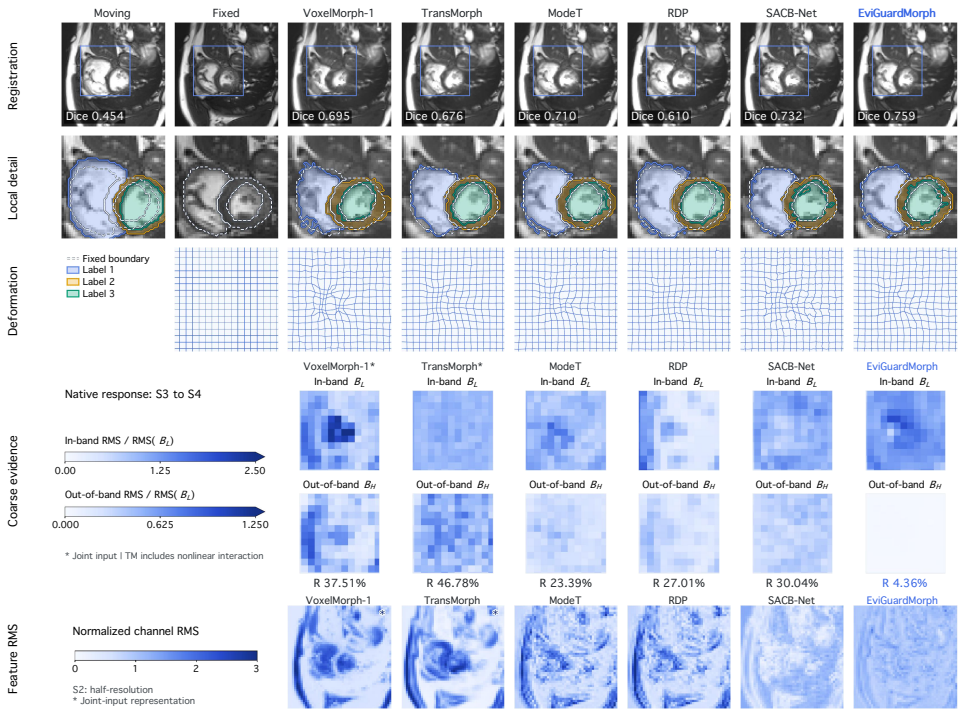


Figure 9: Example registration and feature-response comparison on ACDC.

Scale-wise correction analysis. Fig. 3(c) relates BPBM correction magnitude to the immediate Dice gain at each scale. The additional correction is $\delta_s = \Delta\varphi_s - \varphi^s_{\text{base}}$ for $s \in \{2, 3, 4, 5\}$. Its magnitude is

$$m_s = \frac{1}{|\Omega_f^+|} \sum_{x \in \Omega_f^+} \left\| \left[\mathcal{U}^{s-1}(\delta_s) \right](x) \right\|_2 , \tag{21}$$

where Ω_f^+ contains the non-background voxels of the fixed segmentation. The operator $\mathcal{U}^{s-1}$ repeatedly applies the displacement upsampling and rescaling defined in Eq. (2), expressing the correction in input-grid voxel units. The vector norm is computed after interpolation and then averaged over Ω_f^+.

To measure the correction's contribution, we hold the upstream accumulated displacement fixed and compare its composition with φ^s_{base} against its composition with $\Delta\varphi_s$. At S5, no upstream composition is needed. Both resulting fields are lifted to the input grid and used to warp the original moving segmentation with nearest-neighbor interpolation. Dice is evaluated against the fixed segmentation using the dataset-specific metric. The local gain is $100(\text{Dice}_{\text{corr}} - \text{Dice}_{\text{base}})$ in percentage points. Here, local refers to the current-scale correction, not a spatial crop or an adjacent-scale comparison. Finer-scale matching is not rerun, so these gains describe immediate effects rather than additive contributions to the final Dice. Each point represents one registration pair at one scale.

Qualitative comparisons. Figs. 7–10 compare EviGuardMorph with VoxelMorph-1 and TransMorph as representative convolutional and Transformer-based registration methods, alongside three feature-pyramid methods, ModeT, RDP, and SACB-Net. The visualizations connect registration outcomes with coarse-scale downsampling responses and fine-scale features. Colored contours distinguish anatomical labels, while dashed contours indicate fixed-image boundaries. Across the four examples, EviGuardMorph achieves the highest Dice among the compared methods, with the enlarged views showing improved anatomical contour agreement.

At the S3-to-S4 transition, EviGuardMorph exhibits lower out-of-band-to-in-band response ratios while retaining spatially structured in-band responses. Relative to SACB-Net, the ratios decrease from 7.19% to 1.36% on IXI, from 6.07% to 1.51% on LPBA, from 30.04% to 4.36% on ACDC,

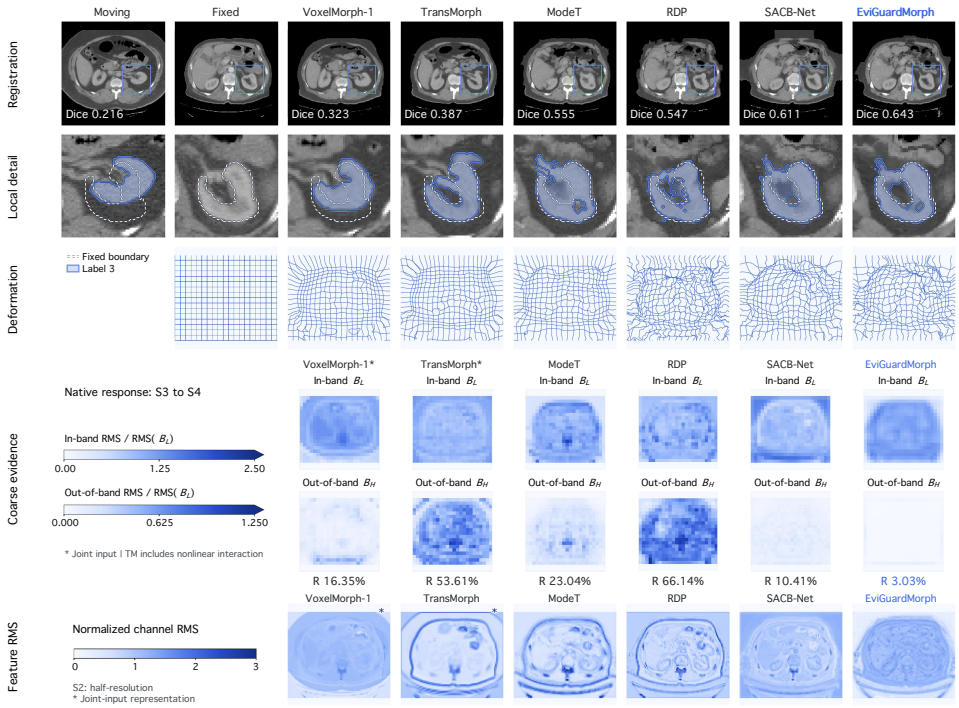


Figure 10: Example registration and feature-response comparison on AbdomenCT.

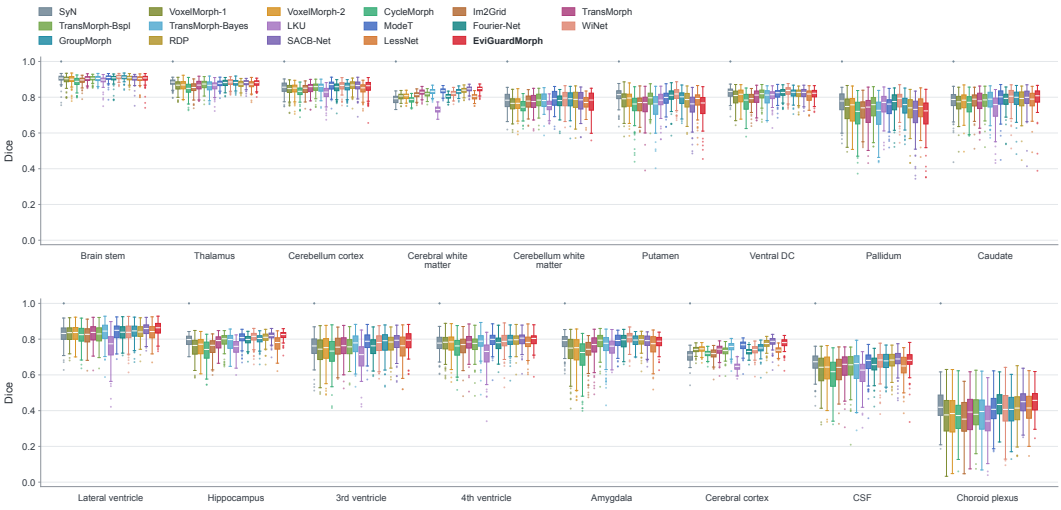


Figure 11: Structure-wise Dice distributions on IXI. Scores for bilateral structures are averaged within each registration pair. Boxes span the first to third quartiles, with internal lines indicating medians. Whiskers extend to the most extreme observations within 1.5 interquartile ranges of the box, and points beyond the whiskers indicate outliers. Higher Dice scores indicate better anatomical alignment.

and from 10.41% to 3.03% on AbdomenCT. These ratios indicate a smaller relative contribution of above-Nyquist inputs to the downsampling output, consistent with FBE's objective of limiting coarse-scale out-of-band interference. This is relevant to pyramid-based registration because coarse displacement estimates pre-align finer-scale features, allowing errors in coarse correspondence estimation to affect subsequent refinement. Bandwidth control must also retain the spatial cues needed for precise localization. In the S2 feature maps, anatomical contours and local variations remain visible despite the weaker relative out-of-band responses at the coarser transition. These observa-



Figure 12: Structure-wise Dice distributions on LPBA. The 54 anatomical labels are grouped into 27 bilateral regions by averaging left and right scores within each registration pair. Boxes span the first to third quartiles, with internal lines indicating medians. Whiskers extend to the most extreme observations within 1.5 interquartile ranges of the box, and points beyond the whiskers indicate outliers. Higher Dice scores indicate better anatomical alignment.

tions are consistent with FBE's scale-dependent design: the structure stream controls bandwidth across the pyramid, while the detail stream and its fusion with structure evidence retain fine-scale localization cues. Their coexistence with improved contour alignment supports the joint objective of suppressing coarse-scale interference and preserving fine-scale spatial information. Together with the controlled filtering ablations in Appendix A.2, the visualizations support integrating prefiltering with structure–detail evidence construction rather than relying on uniform smoothing.

A.4 Structure-wise Registration Performance

To characterize registration performance across anatomical regions, we complement the overall metrics with structure-wise Dice distributions over the test registration pairs, as shown in Figs. 11–14. For bilateral structures in IXI, left and right Dice scores are computed separately and then averaged within each registration pair. Similarly, the 54 anatomical labels in LPBA are summarized as 27 bilateral regions by averaging their left and right Dice scores. ACDC reports results for the right ventricle, myocardium, and left ventricle, while AbdomenCT reports results for 13 abdominal structures. The boxplots summarize the median, interquartile range, and variability across registration pairs for each structure.

Brain structures. Figs. 11 and 12 show that EviGuardMorph maintains relatively high Dice scores across multiple brain structures. On IXI, favorable performance is observed in the lateral ventricles, hippocampus, and cerebral cortex. On LPBA, EviGuardMorph also performs well in the postcentral gyrus and several temporal regions, indicating that its gains extend across different anatomical regions. Performance nevertheless varies across structures. For example, the choroid plexus in IXI exhibits relatively low Dice scores and a wide distribution, highlighting a challenging local structure not fully reflected by the overall average results.

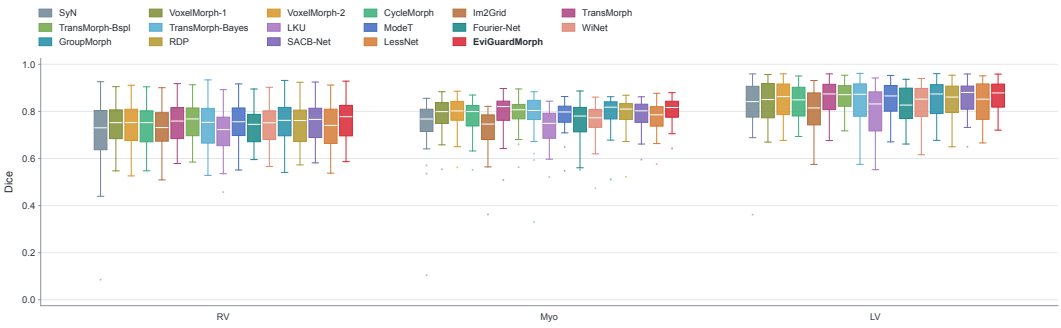


Figure 13: Structure-wise Dice distributions on ACDC for the right ventricle (RV), myocardium (Myo), and left ventricle (LV). Boxes span the first to third quartiles, with internal lines indicating medians. Whiskers extend to the most extreme observations within 1.5 interquartile ranges of the box, and points beyond the whiskers indicate outliers. Higher Dice scores indicate better anatomical alignment.

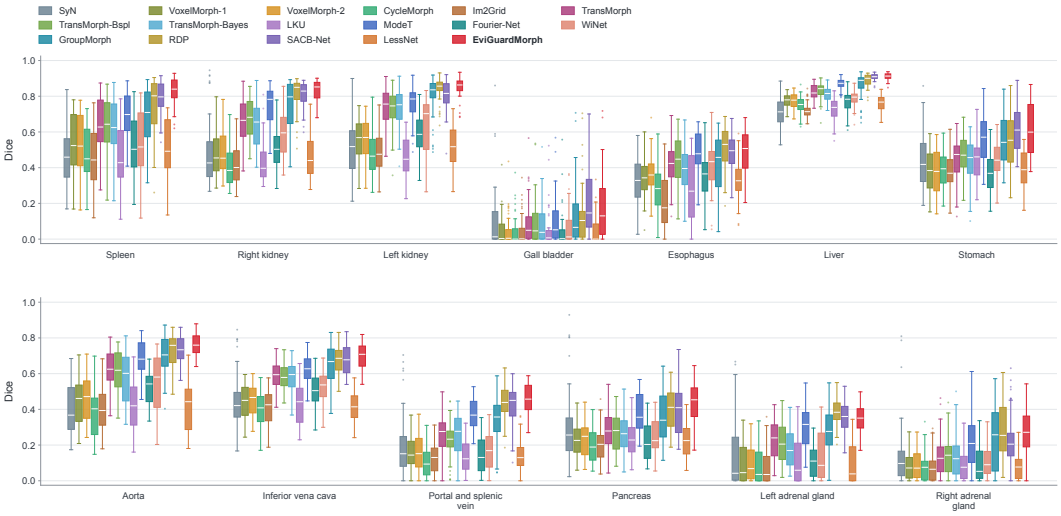


Figure 14: Organ-wise Dice distributions across 13 abdominal structures on AbdomenCT. Boxes span the first to third quartiles, with internal lines indicating medians. Whiskers extend to the most extreme observations within 1.5 interquartile ranges of the box, and points beyond the whiskers indicate outliers. Higher Dice scores indicate better anatomical alignment.

Cardiac structures. Fig. 13 presents Dice distributions for the three cardiac structures in cross-phase registration. EviGuardMorph maintains relatively high Dice scores for the right ventricle, myocardium, and left ventricle, with results across these structures supporting its overall performance. Across methods, Dice scores are generally higher for the left ventricle, whereas the right ventricle exhibits greater variability across registration pairs. These distributions reflect differences in cross-phase registration performance among cardiac structures.

Abdominal structures. As shown in Fig. 14, EviGuardMorph achieves relatively high Dice scores for the liver, spleen, both kidneys, and aorta, while also performing well in the pancreas and the portal and splenic vein region. Performance varies considerably across abdominal structures. Dice scores remain relatively low for the gallbladder and adrenal glands, with considerable variability across registration pairs. These structure-wise results show that the registration advantages of EviGuardMorph extend across multiple abdominal organs while revealing differences in difficulty among structures in multi-organ registration.

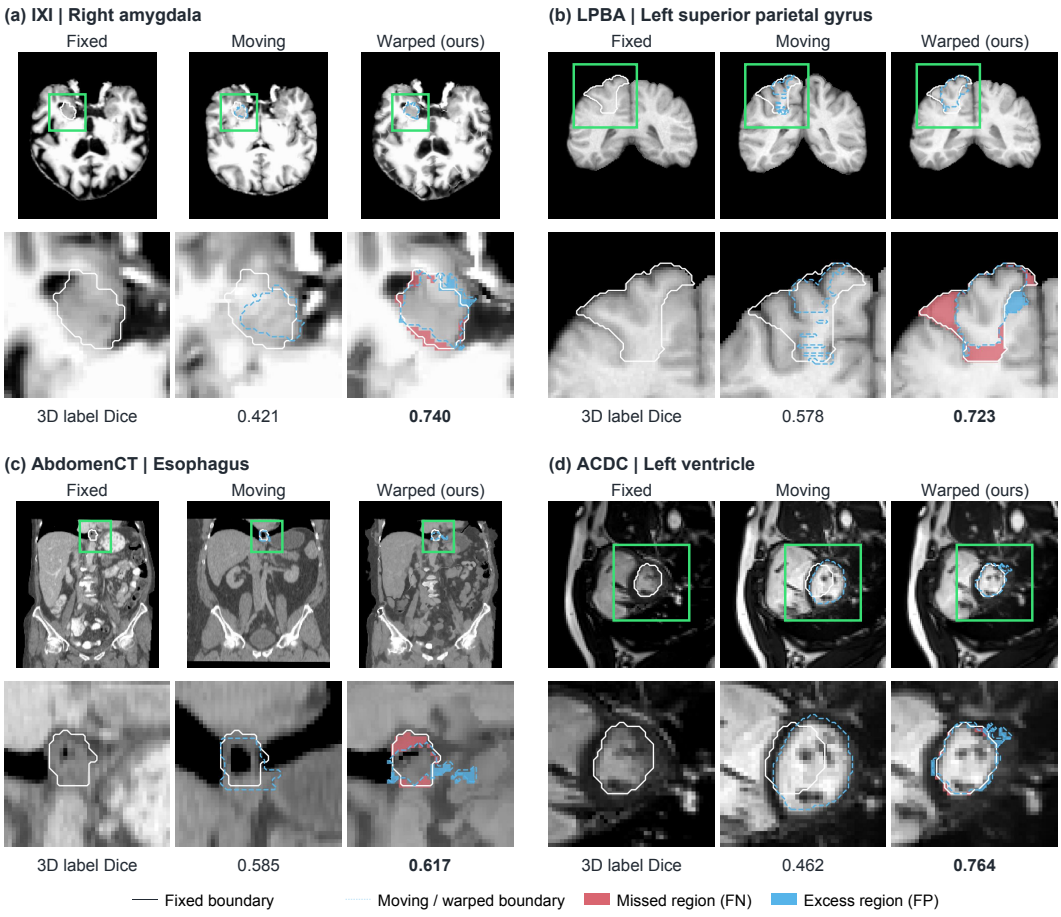


Figure 15: **Residual anatomical mismatch across four datasets.** (a) IXI, (b) LPBA, (c) AbdomenCT, and (d) ACDC. Each panel shows fixed, moving, and warped images above enlargements of the green-boxed ROIs. White solid and cyan dashed contours denote fixed and moving/warped label boundaries, respectively. Coral and cyan fills indicate missed (FN) and excess (FP) label regions after warping. Dice is computed per structure over the full 3D volume.

Overall, the structure-wise analysis is consistent with the aggregate evaluation in the main text and further characterizes the registration performance of EviGuardMorph across diverse brain regions, cardiac tissues, and abdominal organs.

A.5 Limitations Analysis

Fig. 15 presents examples of partial registration failure across four datasets. Although registration improves volumetric overlap, local boundary discrepancies and missed or excess label regions remain. For each dataset, fixed, moving, and warped images are shown with enlarged views of the same ROI to highlight these residual anatomical mismatches.

On IXI, the right amygdala improves from 0.421 to 0.740. Its small spatial extent makes overlap sensitive to local boundary displacement, which may explain the remaining mismatch despite improved localization. On LPBA, the left superior parietal gyrus improves from 0.578 to 0.723, but missed and excess regions remain along its folded boundary. Similar appearances in neighboring cortical regions may make these local correspondences ambiguous. The esophagus in AbdomenCT shows a smaller improvement, from 0.585 to 0.617. Its narrow geometry and inter-subject variation in surrounding anatomy may complicate localization within the local matching window. On ACDC, the left ventricle improves from 0.462 to 0.764, while phase-dependent changes in shape and appearance may contribute to residual contour discrepancies.

These observations highlight a distinction between controlling correspondence reliability and establishing anatomical correctness. FBE reduces out-of-band interference, but does not explicitly enforce anatomical boundary correspondence. BPBM weights corrections using matching entropy and cycle consistency, neither of which guarantees that a candidate matches the correct structure. A locally concentrated, cycle-consistent match can still be anatomically inaccurate. Moreover, finite search windows and reliability-weighted corrections may leave some pre-alignment errors only partially resolved. These mechanisms provide plausible explanations for the residual mismatches and motivate stronger boundary-aware correspondence modeling.