SIGMI: SIGMOID CONTRASTIVE MODELS ESTIMATE POINTWISE MUTUAL INFORMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Histogram-based approximations of mutual information (MI) have long been used as multimodal similarity measures for medical image registration. Representation learning emerged as an alternative approach, but its connection to mutual information is lost when applied to image registration. We connect both lines of work by showing that sigmoid contrastive models like SigLIP estimate pointwise mutual information (PMI): being trained to distinguish samples from a joint distribution from deranged pairs approximating the product of its marginals, their optimal logits recover PMI up to a correction for the sampling class prior. These *sigmoidal mutual information estimators* (SigMIs) can estimate calibrated PMI, independent of the training prior. Their estimation error can be described in terms of excess classification risk, and they provide information-theoretic interpretations of SigLIP's temperature and bias parameters. Applying our findings to medical image registration, we propose fully convolutional SigMIs that learn dense image representations whose dot products estimate local PMI. We train our models on domain-randomized synthetic images generated from just five brain segmentation maps. Across three registration algorithms and four datasets, SigMIs outperform other similarities in 8 of 12 settings. Our code and weights are available here.

1 INTRODUCTION

Multimodal image registration requires a measure of whether two observations are spatially aligned. For three decades, the criterion of choice has been mutual information (MI) (Viola & Wells III, 1997; Maes et al., 1997), which makes no assumption of a functional relation between intensities. MI is the expectation under the joint distribution of the *pointwise* mutual information (PMI),

$$I[U;V] = \mathbb{E}_{(u,v)\sim p_{UV}}\left[\mathrm{PMI}(u,v)\right], \qquad \mathrm{PMI}(u,v) = \log \frac{p_{UV}(u,v)}{p_U(u)\,p_V(v)}, \tag{1}$$

and a registration solver typically requires this pointwise term at every location rather than the global MI estimate to handle a spatially varying joint distribution. Localized MI estimators do provide this pointwise MI but under an additional assumption. Histogram-based MI estimators (Klein et al., 2008; Loeckx et al., 2009; Rivaz et al., 2013) spread a fixed number of samples over an increasing number of bins as spatial degrees of freedom are added, trading bias for variance. Regional MI (Russakoff et al., 2004) avoids binning by estimating the entropies of neighborhood vectors under a Gaussian approximation, at the cost of the distribution-free property that initially motivates the use of MI. Neural estimators leverage a trained network instead of histograms but remain global and must be refitted for each image pair (Nan et al., 2020; Snaauw et al., 2022).

A second line of MI-based registration methods learns modality-invariant descriptors and uses their similarity as the registration criterion (Pielawski et al., 2020; Dey et al., 2025). Both use InfoNCE-like objectives, whose standard form lower bounds MI (Oord et al., 2018). Evaluating that bound requires normalizing the score of a pair against others in its batch, and its value is upper-bounded by $\log B$, where B is the batch size (Poole et al., 2019). At registration time, however, pairs are scored without a batch, i.e., one at a time. The bound thus plays no role in the registration criterion. Outside registration, a popular alternative to InfoNCE is SigLIP (Zhai et al., 2023). By applying an independent binary decision to each score, i.e. without normalizing over the batch, the sigmoid loss allows scaling up the batch size, while also performing better at smaller batch sizes. However, this strategy is usually interpreted as trading the information-theoretic interpretation for scalability.

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.

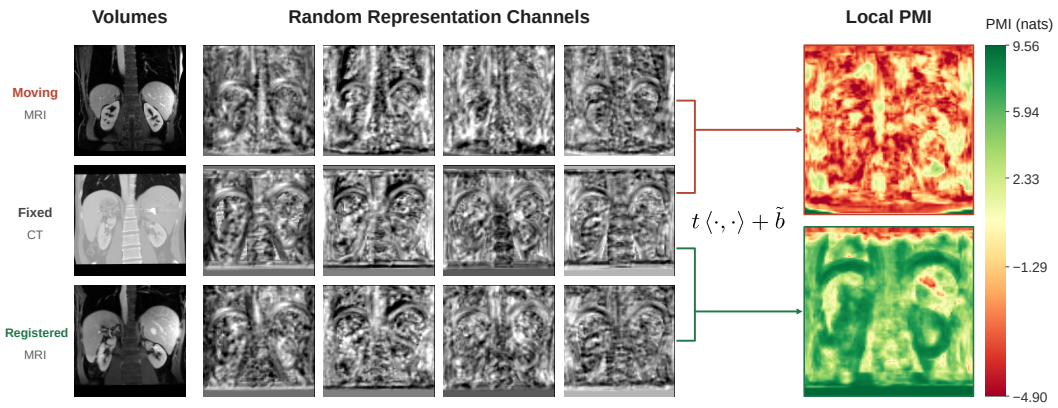

Figure 1: **Estimator for medical images.** Our method extracts dense embeddings for both images. Each embedding depends on a cubic local image window centered on it. All local PMI estimates are computed with element-wise dot products of the two embedding fields.

In this work, we show that the sigmoid objective retains an information-theoretic interpretation: an encoder trained with the SigLIP objective is already a pointwise mutual information estimator and can be leveraged for image registration. Subtracting the log odds of the implied class prior $\pi = 1/B$ from that score, where B is the training batch size, yields an estimate of PMI under no assumption of class balance. This positions SigLIP in the two-step family of MI estimators, in which the MI estimate is obtained from the trained discriminator rather than from the value of its objective (Liao et al., 2020; Letizia et al., 2024). Since that score is affine in the cosine similarity between unit embeddings φ_θ, with temperature t and bias b,

$$\widehat{\mathrm{PMI}}(u, v) = t\left\langle \varphi_\theta(u), \varphi_\theta(v) \right\rangle + \tilde{b}, \qquad \tilde{b} = b - \log \frac{\pi}{1-\pi}, \tag{2}$$

both coefficients can be interpreted: $2t$ bounds the range of the estimator in nats, and $\tilde{b}$ is its center.

Contributions. We (i) show that the SigLIP objective is a two-step MI estimator that recovers pointwise MI under *any* class prior; (ii) give its closed form and a bound on its pointwise error; (iii) instantiate it for multimodal medical image registration as a fully 3D convolutional encoder that outputs a dense, solver-agnostic PMI estimator, trained only on domain-randomized synthesis from five anatomical templates and no real images; and (iv) evaluate it exhaustively across four datasets and three solver classes, including MRI, CT and ultrasound modalities and brain and abdominal anatomy, showing that SigMI improves registration accuracy for 8 of 12 settings over handcrafted and learned measures. Plugged into existing solvers, our method is also competitive with end-to-end methods that design solver and measure together.

2 Related work

Handcrafted similarity measures. Alternatives to MI (Viola & Wells III, 1997; Maes et al., 1997) trade its generality for an assumption on the intensity relation. The correlation ratio (Roche et al., 1998) and its local variants (Gong et al., 2018; Honkamaa & Marttinen, 2025) assume that one intensity is a function of the other, globally or within a window. LC^2 (Wein et al., 2013) models this relation locally as a linear combination of the other intensity and its gradient magnitude. MIND-SSC (Heinrich et al., 2012; 2013) assumes that local self-similarity is shared across modalities.

Learned similarity measures and representations. Learned measures relax these assumptions, but their values have no information-theoretic interpretation, as their scale is set by a teacher metric or a training loss. Most of them are also trained for specific modalities, and none is evaluated on a modality pair unseen during training: Simonovsky et al. (2016) classify aligned versus misaligned patches from registered T1-T2 pairs, CoMIR (Pielawski et al., 2020) trains one network per modality on aligned images, and ContraReg (Dey et al., 2022) and Mok et al. (2024) are trained on images of the modalities they register. DISA (Ronchetti et al., 2023) is trained once, on a specific modality pair,

and transfers to unseen anatomies and modality pairs, but its feature dot product aims to approximate LC^2 and therefore keeps its local linear model. Closest to our setting, Anatomix (Dey et al., 2025) is also trained contrastively on domain-randomized synthetic volumes, but its positives are all voxels sharing a synthetic label (Khosla et al., 2020) and not corresponding voxels. Their features thus tend to be homogeneous within regions and well-suited for segmentation tasks. However, their best registration performance relies on the concatenation of their features with MIND-SSC.

Neural estimation of mutual information. Variational MI estimators (Belghazi et al., 2018; Poole et al., 2019) optimize a lower bound whose variance can grow exponentially with the true MI (Song & Ermon, 2019). They have been used for image registration to estimate a global MI value per image pair (Nan et al., 2020; Snaauw et al., 2022), with DRMIME (Nan et al., 2020) retraining its estimator for every pair. Two-step estimators (Liao et al., 2020; Letizia et al., 2024) train a discriminator to separate joint from product of marginals samples and then recover the density ratio from its output, which after removing the class prior is a pointwise estimate (Tsai et al., 2020). These estimators are fitted to one joint distribution and must be retrained whenever it changes (Dorent et al., 2025). In word embeddings, Levy & Goldberg (2014) showed that skip-gram with k negative samples factorizes a PMI matrix shifted by $\log k$, similar to our prior correction.

3 METHOD

We construct SigMI in two steps. We first derive a PMI estimator from the prior-corrected SigLIP score, and show that it recovers PMI under any class prior, as described in sections 3.2 to 3.4. We then turn this estimator into a dense, solver-agnostic similarity measure for 3D multimodal image registration, trained without any real images, as described in section 3.5.

3.1 PRELIMINARIES: THE SIGLIP OBJECTIVE

Sigmoid contrastive learning on paired data. SigLIP (Zhai et al., 2023; Tschannen et al., 2025) learns representations from multimodal paired data, such as images and their captions. Two encoders φ_θ and ψ_θ map a batch of B pairs (u_i, v_i) to unit-norm embeddings $h_i = \varphi_\theta(u_i)$ and $g_i = \psi_\theta(v_i)$. Every combination (i, j) in the batch is scored by an affine function of the cosine similarity,

$$s_{ij} \;=\; t\,\langle h_i, g_j\rangle + b, \tag{3}$$

with a learned temperature $t > 0$ and a learned bias b. During training, scores of the B matched pairs ($i = j$) are raised, while the $B(B-1)$ mismatched ones ($i \neq j$) are lowered. The InfoNCE objective (Oord et al., 2018), used for instance by CLIP (Radford et al., 2021), normalizes the unbiased scores over the batch with a softmax. SigLIP instead treats each score as an independent binary decision, positive for paired elements and negative otherwise:

$$\mathcal{L}_{\text{SigLIP}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\Big[y_{ij}\log\sigma(s_{ij}) + (1-y_{ij})\log\big(1-\sigma(s_{ij})\big)\Big], \qquad y_{ij} = \mathbf{1}[i=j]. \tag{4}$$

Since this loss is a sum over pairs, no normalization over the batch is needed. Zhai et al. (2023) use this to reduce memory and scale the batch size, while outperforming CLIP at smaller batch sizes.

Temperature and bias. The learnable bias b has no counterpart in InfoNCE. Zhai et al. (2023) found that without it, the early optimization steps aim to compensate for the heavy imbalance from the many negatives. They also found that the optimized bias value increases as the batch size decreases. The objective, score and scalars above are SigLIP; the rest of this section is our construction.

3.2 THE SIGLIP OBJECTIVE IS A CLASSIFICATION RISK UNDER AN UNBALANCED PRIOR

Let U and V be random variables with joint distribution p_{UV} and product of marginals $p_U \otimes p_V$. Let $\text{PMI} \triangleq \log \mathrm{d}p_{UV}/\mathrm{d}p_U \otimes p_V$ be the *pointwise mutual information* (PMI), whose expectation under p_{UV} is the mutual information $I[U;V]$. Following Dorent et al. (2025), consider the mixture model:

$$(U,V) \mid Z{=}1 \sim p_{UV}, \qquad (U,V) \mid Z{=}0 \sim p_U \otimes p_V, \qquad Z \sim \mathcal{B}(\pi), \tag{5}$$

with marginal $m_\pi = \pi p_{UV} + (1-\pi)p_U \otimes p_V$. The population risk of the discriminator $\sigma(s)$ is:

$$\mathcal{L}_\pi(s) \;\triangleq\; -\pi \mathop{\mathbb{E}}_{p_{UV}}\big[\log\sigma\big(s(U,V)\big)\big] \;-\; (1-\pi)\mathop{\mathbb{E}}_{p_U\otimes p_V}\big[\log\big(1-\sigma(s(U,V))\big)\big]. \tag{6}$$

Proofs of all results in this section are given in appendix A.

Proposition 1 (Identification)**.** *Let the B pairs of a batch be i.i.d. draws from p_{UV}. Then $\mathbb{E}\big[\mathcal{L}_{\mathrm{SigLIP}}\big] = B\,\mathcal{L}_\pi(s)$ with $\pi = 1/B$. More generally, weighting diagonal terms by π/B and off-diagonal terms by $(1-\pi)/\big(B(B-1)\big)$ yields $\mathcal{L}_\pi(s)$ for any $\pi \in (0,1)$ at fixed B.*

The batch size thus determines the class prior, $\pi = 1/B$. The objective is unbalanced by construction, and severely at typical batch sizes. Because negatives are restricted to $i \neq j$, the negative sampling method for $p_U \otimes p_V$ corresponds to the derangement strategy from (Letizia et al., 2024) and avoids the fixed points that a random permutation would leave.

3.3 SigMI: a pointwise MI estimator from the prior-corrected logit

We now build a PMI estimator from a trained SigLIP score.

Theorem 2 (Bayes-optimal score)**.** *For every $\pi \in (0,1)$, the minimizer of $\mathcal{L}_\pi$ over measurable scores is, m_π-almost everywhere,*

$$s^\star_\pi(u,v) \;=\; \mathrm{PMI}(u,v) \;+\; \log\frac{\pi}{1-\pi}. \tag{7}$$

Our PMI estimator follows by evaluating that logit and removing the prior: for any score s trained with the SigLIP objective, we define the SigMI estimate

$$\widehat{\mathrm{PMI}}(u,v) \;\triangleq\; s(u,v) - \log\frac{\pi}{1-\pi}. \tag{8}$$

Corollary 3 (The training prior cancels)**.** $\widehat{\mathrm{PMI}}$ *is exact at the optimum for every prior: for any $\pi_1, \pi_2 \in (0,1)$, $s^\star_{\pi_1} - \log\frac{\pi_1}{1-\pi_1} = s^\star_{\pi_2} - \log\frac{\pi_2}{1-\pi_2} = \mathrm{PMI}$.*

Substituting the SigLIP score equation (3) into equation (8) gives the form we actually compute.

Corollary 4 (PMI is a shifted, temperature-scaled dot product)**.** *For the score equation* (3)*,*

$$\widehat{\mathrm{PMI}}(u,v) \;=\; t\,\big\langle \varphi_\theta(u),\, \psi_\theta(v)\big\rangle \;+\; \tilde{b}, \qquad \tilde{b} \;\triangleq\; b - \log\frac{\pi}{1-\pi}. \tag{9}$$

Range and offset. The temperature and the bias are then determined quantities rather than free parameters. Since $\langle \varphi_\theta(u), \psi_\theta(v)\rangle \in [-1,1]$, the estimates lie in $[\tilde{b}-t, \tilde{b}+t]$, an interval of width $2t$ nats centered on $\tilde{b}$. This interval bounds the representable range. It is attained only if the embeddings use the full cosine range, which the optima of the sigmoid loss need not do (Lee et al., 2024). Writing c and r for the center and half-width of the PMI range $[\mathrm{PMI}_{\min}, \mathrm{PMI}_{\max}]$ to be represented, clipping can only be avoided if

$$t \;\geq\; r \qquad \text{and} \qquad |\tilde{b} - c| \;\leq\; t - r. \tag{10}$$

At the smallest admissible temperature, $t = r$, the offset must equal c, which is nonzero in general: the expected value of PMI is $I[U;V] \geq 0$ under p_{UV} and $-D_{\mathrm{KL}}(p_U \otimes p_V \,\|\, p_{UV}) \leq 0$ under $p_U \otimes p_V$, and nothing makes its range symmetric about zero. Fixing b at the prior log odds, i.e. $\tilde{b} = 0$, would require $t \geq r + |c|$, which is why b must be learned, as in SigLIP. Both t and $\tilde{b}$ are read off the trained head, and estimates that accumulate at the edges of the used range signal clipping.

The bias as part of the estimate. Zhai et al. (2023) initialize $b = -10$ to start training close to the class prior. By theorem 2 the trained score is PMI plus $\log\frac{\pi}{1-\pi}$, a constant set by the batch construction and not by the data, which equation (8) removes whatever value b takes, leaving the offset $\tilde{b}$. Proposition 1 gives that constant as $-\log(B-1)$, which -10 matches only for batches of order 10^4; we initialize b such that $\tilde{b} = 0$ at the start of training.

3.4 SigMI at finite capacity: excess risk controls the pointwise error

Theorem 2 describes the optimum, which a trained score s only approaches. Let $\varepsilon = s - s^\star_\pi$ denote its logit error and $\Delta_\pi(s) \triangleq \mathcal{L}_\pi(s) - \mathcal{L}_\pi(s^\star_\pi) \geq 0$ its excess risk. Since equation (8) shifts s and $s^\star_\pi$ by the same constant, $\widehat{\mathrm{PMI}} - \mathrm{PMI} = \varepsilon$. Bounding the error of the PMI estimate thus simplifies to bounding ε by $\Delta_\pi(s)$, which training minimizes.

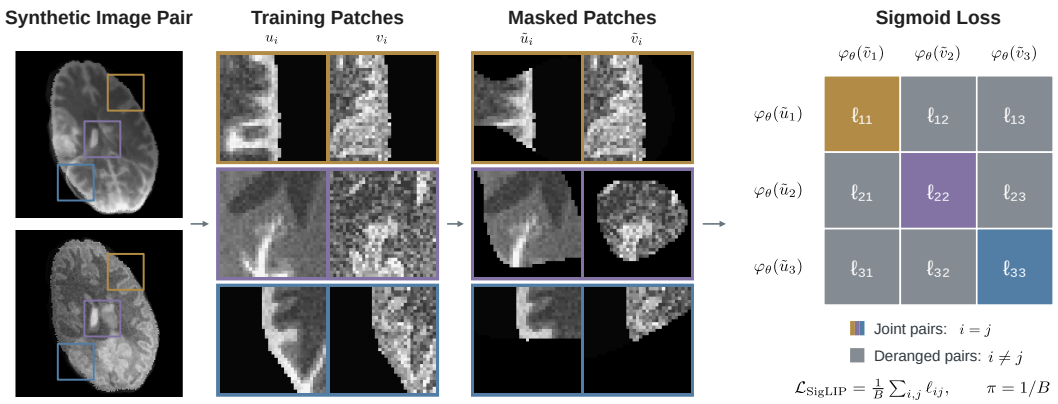

Figure 2: **Training pipeline.** Two independently rendered contrasts of the same synthetic brain provide matched 3D patches. Independently applied ellipsoidal masks break the correspondence shortcut via foreground silhouette. Diagonal elements of the sigmoid loss matrix are joint pairs, and off-diagonal elements form the deranged negative pairing required by Letizia et al. (2024).

Theorem 5 (Excess risk is a Bregman divergence in the logit)**.** *With* $A(z) = \log(1 + e^z)$,

$$\Delta_\pi(s) \;=\; \mathop{\mathbb{E}}_{m_\pi}\left[A\big(s^\star_\pi + \varepsilon\big) - A\big(s^\star_\pi\big) - \varepsilon\,\sigma\big(s^\star_\pi\big)\right] \;=\; \tfrac{1}{2}\mathop{\mathbb{E}}_{m_\pi}\left[\sigma'(\xi)\,\varepsilon^2\right] \tag{11}$$

for some ξ *between* $s^\star_\pi$ *and* s*. If moreover* $|s|, |s^\star_\pi| \le M$*, then with* $c_M = \sigma(M)\big(1 - \sigma(M)\big)$,

$$\tfrac{c_M}{2}\mathop{\mathbb{E}}_{m_\pi}\left[\varepsilon^2\right] \;\le\; \Delta_\pi(s) \;\le\; \tfrac{1}{8}\mathop{\mathbb{E}}_{m_\pi}\left[\varepsilon^2\right]. \tag{12}$$

Corollary 6 (Error under the joint, and the price of imbalance)**.** *Under the hypotheses of theorem 5,*

$$\mathop{\mathbb{E}}_{p_{UV}}\left[\big(\widehat{\mathrm{PMI}} - \mathrm{PMI}\big)^2\right] \;\le\; \frac{1}{\pi}\mathop{\mathbb{E}}_{m_\pi}\left[\varepsilon^2\right] \;\le\; \frac{2\,\Delta_\pi(s)}{c_M\,\pi}. \tag{13}$$

The estimate is therefore consistent: as the excess risk vanishes, $\widehat{\mathrm{PMI}}$ converges to PMI in mean square under the joint, with no assumption on class balance. The guarantee weakens as $\pi \to 0$, the price of training with one positive per batch, which section 4.1 tests directly.

3.5 SigMI as a dense similarity measure for 3D multimodal registration

We define the joint distribution over pairs of co-located patches (U, V) from two images whose underlying anatomy is perfectly aligned. Non-overlapping patches from different locations approximate the product of marginals, such that a batch of B patch pairs trains the estimator under $\pi = 1/B$.

Synthetic training data. Sampling this joint distribution requires perfectly aligned multimodal images. Such pairs are rarely available for real data, since obtaining them requires solving the registration problem. An estimator trained on a real dataset would also be limited to the modalities and anatomies it contains. We instead train it on synthetic data only, following domain-randomized synthesis (Hoffmann et al., 2021; Billot et al., 2023; Dey et al., 2024; 2025; Valabregue et al., 2026), which provides perfect alignment by construction and contrast agnosticism. Using the generative model of Valabregue et al. (2026), only five brain label templates are upsampled to 0.25 mm, strongly augmented by random erosion, dilation, affine and elastic deformations, and rendered at 0.75 mm with two independent random contrasts. Each training pair is then resampled to a voxel spacing drawn uniformly between 1.5 and 3.0 mm, such that fixed-size patches cover anatomy at varying scales, contain heterogeneous intensities, and remain abstract enough to transfer beyond brain anatomy. Spatial augmentations are shared within a patch pair to preserve correspondence, and appearance augmentations are drawn independently per patch. Finally, trivially separable joint and product of marginals distributions yield strong classifiers but poor density-ratio estimates (Rhodes

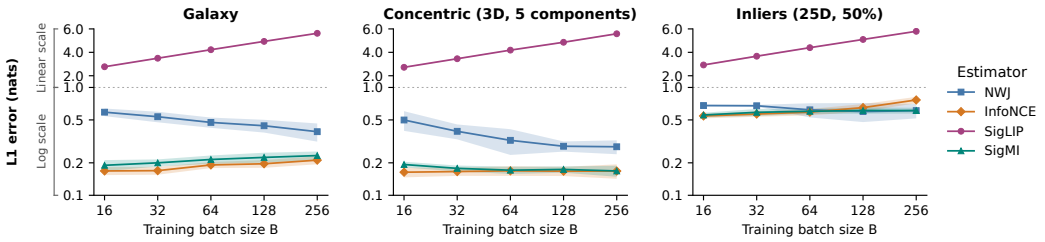

Figure 3: **SigMI is a competitive PMI estimator across training batch sizes.** Pairwise L_1 error on three BMI v2 tasks, as mean and standard deviation over seeds. The vertical axis is logarithmic below 1 nat and linear above. Without the prior correction, SigLIP is offset by $\log(B-1)$.

et al., 2020). Since a large fraction of patches partially contains background, the foreground silhouette alone would often reveal correspondence. We therefore randomly mask each patch with an independent, randomly deformed ellipsoidal field of view, forcing the estimator to rely on visible anatomy rather than silhouettes. The pipeline is illustrated in Figure 2 and detailed in appendix B.1.

Shared, fully convolutional 3D encoder. A shared encoder maps each patch to an L_2-normalized embedding, and local PMI is estimated from the dot product of two embeddings with equation (9). The encoder is fully convolutional, and each embedding depends only on its receptive field, a cubic window centered at its location. We can therefore train it on patches of the size of its receptive field and apply it to whole images of arbitrary size, where a single forward pass returns a dense field of the same embeddings. Each batch of $B = 32$ patch pairs combines patches from several image pairs. We instantiate SigMI-R9 and SigMI-R17 with receptive fields of 9 and 17 voxels. For registration, we evaluate local PMI at multiple locations and use its negative mean as the loss. Details on model architectures, optimization settings, and the application to registration are given in appendix B.2.

4 Experiments

Local PMI estimates from SigMI before and after registration are visualized in Figure 4. Below, we present experiments that evaluate SigMI as a pointwise MI estimator against ground truth (section 4.1), as a similarity measure across anatomies, modalities, and registration solvers (section 4.2), and against complete registration methods (section 4.3).

4.1 SigMI as a mutual information estimator

Setup. This experiment tests the claim of section 3.3: once corrected for the class prior, the SigLIP score estimates PMI whatever the training batch size. We use the 13 Bend and Mix Model tasks of the BMI v2 benchmark (Czyż et al., 2025), whose ground-truth PMI is known for each pair, with 2,500 training and 2,500 test pairs per task. *SigLIP* is an encoder-based critic trained with the sigmoid loss: two MLP encoders map U and V to unit-norm embeddings scored by equation (3). *SigMI* is the same model with the prior correction of equation (8). We compare it with *NWJ* (Nguyen et al., 2010), whose optimal critic directly gives PMI, and with *InfoNCE* (Oord et al., 2018), trained with the same encoders, whose scores give PMI only after normalization over a reference set, here the test set. We report the pairwise L_1 error of the PMI estimates, averaged over 10 random seeds, for training batch sizes $B \in \{16, 32, 64, 128, 256\}$ (appendix C.1).

Results. Figure 3 shows three representative tasks. Without the correction, the SigLIP score is offset by the log odds of the training prior ($\log(B-1)$). Its error thus grows with the batch size and exceeds 5.5 nats at $B = 256$ on all 13 tasks. With the correction, SigMI has a low error that stays nearly constant across batch sizes. At $B = 256$, its error never exceeds InfoNCE's by more than 0.09 nats on all tasks, and is lower on three. It is lower than that of NWJ on 10 of the 13 tasks. InfoNCE's estimate of a given pair depends on the reference set used to normalize it: with fewer than about 32 reference samples, its error varies by several tenths of a nat across reference

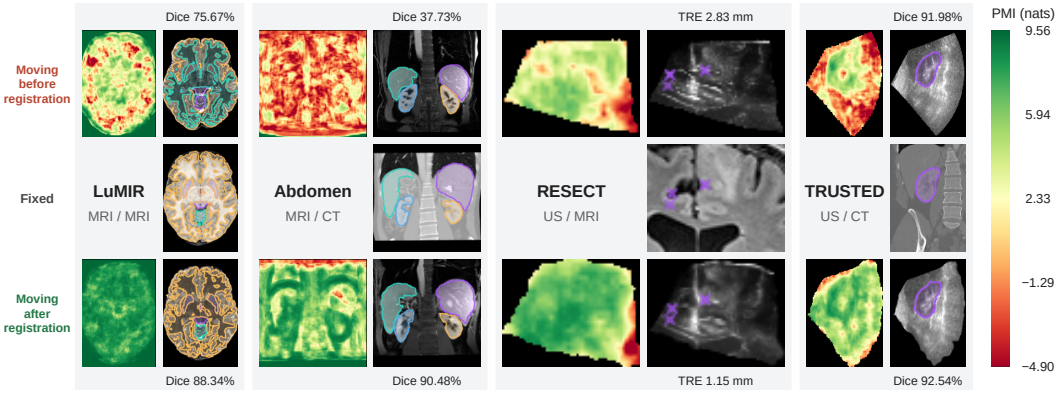

Figure 4: **Qualitative registration examples across four datasets.** Modalities are listed in moving/fixed order. Volume-wise Dice and TRE are reported for each sample before and after SigMI-R17 registration with MultiresSVF. At annotated regions, our calibrated SigMI estimator assigns low initial PMI to datasets with low initial alignment, and moderate PMI to better-aligned datasets.

sets, and it stabilizes only with hundreds of samples (appendix figure 6). SigMI scores each pair on its own and has no such reliance, which is the setting of registration. The full results are reported in appendix C.1.

4.2 Evaluation across anatomies, modalities, and registration solvers

Setup. This experiment tests whether our SigMI estimator trained only on synthetic brain images transfers to real anatomies, modalities, and registration solvers. We use four public benchmarks: *LuMIR* from Learn2Reg 2025 (Chen et al., 2026) (27 brain MRI pairs with T1, high-field T1, and T2 sequences; Dice over 10 structures), *Abdomen* from Learn2Reg 2021 (Hering et al., 2022) (8 T1 MRI to CT pairs; Dice over liver, kidneys, and spleen), *RESECT* (Xiao et al., 2017) (brain FLAIR MRI to intra-operative ultrasound pairs from 22 patients; landmark TRE following the CuRIOUS 2018 protocol (Xiao et al., 2019)), and *TRUSTED* (Ndzimbong et al., 2025) (59 kidney CT to ultrasound pairs from 39 patients; Dice). Together, they cover brain and abdominal anatomy and four cross-modal pairs (T1–T2 MRI, MRI–CT, MRI–ultrasound, and CT–ultrasound). Validation/test splits contain 9/18, 4/4, 8/14, and 12/47 pairs, respectively (details in appendix C.2).

We evaluate each similarity measure with three solvers from widely used classes: *MultiresSVF* (Honkamaa & Marttinen, 2025), a multi-resolution diffeomorphic solver using scaling and squaring (Arsigny et al., 2006); *ConvexAdam* (Siebert et al., 2024), combining convex optimization over a displacement cost volume with Adam instance optimization; and *FireANTs SyN* (Jena et al., 2026), which registers both images to a midpoint space as in ANTs SyN (Avants et al., 2008; Tustison et al., 2021). For each solver and similarity measure, learning rates and regularization weights are tuned on validation pairs; other settings are fixed per dataset and solver (details in appendix C.3).

Both variants, SigMI-R9 and SigMI-R17 (section 3.5), are compared with existing similarity measures from the families of section 2: local intensity dependence with squared LNCC (Avants et al., 2008), which assumes a locally affine relation, and SRWCR (Gong et al., 2018), a local correlation ratio weighted toward the window center; global MI with Mattes MI (Mattes et al., 2001), a histogram-based MI estimate as in Viola & Wells III (1997), and NMI (Studholme et al., 1999), which is less sensitive to partial image overlap; self-similarity with MIND-SSC (Heinrich et al., 2012; 2013); and learned features with Anatomix (Dey et al., 2025), which we use without the MIND-SSC concatenation to isolate its feature contribution (appendix C.4).

Results. Table 1 presents results on the held-out test pairs. SigMI-R9 improves upon the commonly used Mattes MI and NMI in all 12 dataset–solver combinations: relative to Mattes MI, it gains 5 to 20 Dice points on Abdomen, 4.6 to 6.2 on TRUSTED, and 0.3 to 1.5 on LuMIR, and reduces the RESECT TRE by 1.5 to 19.1 mm. Further, a SigMI variant outperforms all baseline sim-

Table 1: **Multimodal multi-organ registration performance.** Entries are mean (population standard deviation) across scored cases. Dice is in percent; TRE is in mm. TRUSTED uses voxel Dice on the prepared CT grid. Bolding and underlining represent the best and second-best results.

		LuMIR	Abdomen	RESECT	TRUSTED
Solver	Similarity	Dice (%) ↑	Dice (%) ↑	TRE (mm) ↓	Dice (%) ↑
	Unregistered	75.33 (3.88)	32.65 (11.39)	5.60 (4.45)	84.74 (3.89)
MultiresSVF	Squared LNCC	<u>88.02</u> (2.19)	66.89 (38.38)	4.60 (4.62)	<u>90.49</u> (2.47)
	Mattes MI	86.93 (3.18)	55.93 (23.14)	5.18 (4.52)	86.19 (3.65)
	NMI	87.09 (3.03)	63.63 (23.19)	5.29 (4.39)	85.91 (3.69)
	SRWCR	87.84 (2.61)	65.79 (34.61)	4.34 (5.09)	88.81 (2.74)
	MIND-SSC	86.68 (2.24)	<u>78.98</u> (15.22)	<u>3.53</u> (3.58)	87.08 (3.59)
	Anatomix	82.60 (3.52)	37.76 (11.19)	16.87 (9.04)	85.09 (3.08)
	SigMI-R9	**88.19** (2.43)	75.93 (14.79)	3.64 (4.41)	**90.79** (2.37)
	SigMI-R17	87.80 (2.37)	**85.49** (4.03)	**3.22** (4.24)	89.32 (2.59)
ConvexAdam	Squared LNCC	**87.97** (2.25)	<u>73.75</u> (16.05)	18.82 (4.03)	78.83 (16.19)
	Mattes MI	86.90 (2.94)	59.24 (18.37)	24.61 (4.34)	78.28 (17.26)
	NMI	86.88 (3.21)	59.36 (20.97)	22.07 (3.86)	78.89 (15.74)
	SRWCR	<u>87.73</u> (2.45)	65.16 (17.59)	22.89 (5.22)	79.93 (13.01)
	MIND-SSC	86.34 (2.42)	**81.72** (9.74)	6.06 (1.90)	18.54 (13.14)
	Anatomix	82.23 (3.92)	40.02 (20.06)	13.80 (5.08)	80.31 (14.09)
	SigMI-R9	87.24 (3.00)	64.60 (25.69)	<u>5.50</u> (1.14)	<u>84.50</u> (4.07)
	SigMI-R17	86.32 (3.35)	60.23 (24.08)	**3.84** (1.22)	**84.51** (4.04)
FireANTs SyN	Squared LNCC	86.68 (2.77)	72.37 (17.67)	4.46 (4.66)	89.16 (2.92)
	Mattes MI	85.73 (3.14)	58.10 (22.92)	4.89 (4.47)	84.77 (3.88)
	NMI	86.42 (3.11)	44.96 (23.35)	4.99 (4.43)	85.83 (3.62)
	SRWCR	86.10 (2.71)	57.12 (26.48)	4.83 (4.63)	87.31 (3.10)
	MIND-SSC	86.60 (2.38)	**81.42** (11.08)	**2.51** (3.08)	86.95 (3.94)
	Anatomix	82.12 (3.80)	40.73 (23.96)	5.25 (4.23)	84.92 (3.59)
	SigMI-R9	**87.18** (3.27)	71.85 (20.42)	<u>3.34</u> (3.63)	<u>89.34</u> (3.12)
	SigMI-R17	<u>86.95</u> (2.78)	<u>80.14</u> (10.00)	3.35 (4.26)	**89.94** (3.11)

ilarity measures in 8 of the 12 combinations and gives the best result on three of the four datasets. In contrast, the strongest baseline changes with the dataset, from squared LNCC on LuMIR and most TRUSTED settings to MIND-SSC on Abdomen and RESECT, where MIND-SSC with FireANTs SyN reaches the best overall TRE of 2.51 mm. Consequently, SigMI-R9 has the best average rank over the 12 combinations (2.3, against 3.3 for squared LNCC and 3.8 for MIND-SSC) and never ranks below fourth. We emphasize that these results are obtained with a single estimator per receptive field, trained without access to any real image. Figure 4 shows qualitative registration examples and local PMI maps before and after registration.

4.3 Comparison to full registration approaches

Setup. We compare our best SigMI configuration from section 4.2, SigMI-R9 on LuMIR and SigMI-R17 on RESECT, with MultiresSVF, against complete registration methods. *multiGradICON* (Demir et al., 2024) is a registration foundation model trained on CT and MRI datasets, which we refine with 50 iterations of instance optimization per pair. *ANTs SyN* (Tustison et al., 2021; Avants et al., 2008) is a widely used multi-resolution, symmetric diffeomorphic method, which we run with Mattes MI. *SynthMorph* (Hoffmann et al., 2021) is a VoxelMorph (Balakrishnan et al., 2019) variant trained on domain-randomized random shapes, i.e., like SigMI, without real images. *LoCoR* (Honkamaa & Marttinen, 2025) was designed with the MultiresSVF solver and approximates a local correlation ratio with neural basis functions fitted to each image pair jointly with the deformation. The hyperparameters of ANTs, SynthMorph, and LoCoR are tuned on the validation pairs

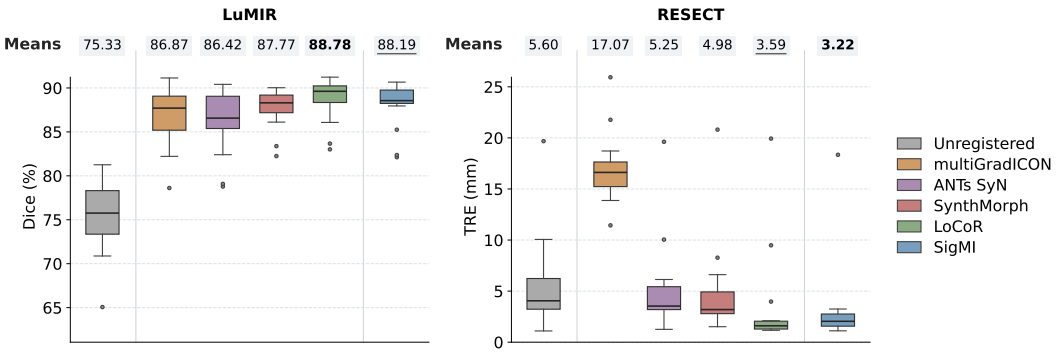

Figure 5: **Registration results for full registration approaches.** Boxes show the interquartile range (25th–75th percentiles), with the median marked by a horizontal line. Whiskers extend to the most extreme observations within 1.5 times the interquartile range. Points indicate observations beyond the whiskers. SigMI shows strong registration performance and low variability on both datasets.

with the protocol of Honkamaa & Marttinen (2025), and multiGradICON uses its default instance optimization (appendix C.5).

Results. Figure 5 presents the results. LoCoR shows the best medians on both datasets, with 89.62% Dice on LuMIR and 1.61 mm TRE on RESECT, followed by SigMI with 89.56% Dice and 2.04 mm TRE, respectively. Since LoCoR shows three negative outliers, SigMI wins RESECT when comparing means. SynthMorph performs similarly on LuMIR, but shows higher variance and reduced performance on RESECT. SigMI consistently shows better performance and lower variance than ANTs SyN and multiGradICON. The latter, trained on CT and MRI only, does not transfer to RESECT, where its mean TRE exceeds that of unregistered images (17.07 vs. 5.60 mm).

5 Discussion and conclusion

Limitations and future work. SigMI is trained once and then used as a fixed PMI estimator. While its consistency holds under the training distribution only (corollary 6), its use on real images assumes that an estimator trained on synthetic pairs remains valid. Experiments in section 4.2 support this assumption on four benchmarks without guaranteeing it for every anatomy or modality. Its range is also fixed at training time: estimates are bounded by $\tilde{b} \pm t$, or more narrowly if the embeddings do not use the full cosine range (Lee et al., 2024). We did not observe any clipping in our experiments; on new data, a lower-bounded temperature such as $t = t_0 + \exp(s)$ could reduce the risk of clipping. Moreover, since patches from a few image pairs are drawn in each batch, the negative distribution is conditional on spacing and subject rather than the exact product of marginals. We argue here that since the estimator should focus on the anatomy and not the contrast nor the scale, this negative sampling approach can be interpreted as a hard negative mining strategy. Finally, our comparison tunes only learning rates and regularization weights for each similarity measure. Although this protocol treats all measures equally, tuning further solver settings for each measure, such as the regularizer type or the ConvexAdam search range, could improve their results.

Conclusion. The sigmoid contrastive objective, usually adopted for its scalability, retains an information-theoretic interpretation: once the log odds of the class prior are removed, the SigLIP score is a pointwise MI estimator under any class prior, whose temperature and bias bound its range. This matters for multimodal registration, which needs a local measure of dependence: SigMI provides one without the additional assumptions of localized MI estimators, and with the information-theoretic meaning that learned similarity measures lack. In practice, trained only on synthetic images from five brain templates, a SigMI variant is the best similarity measure in 8 of 12 dataset-solver combinations, across brain and abdominal anatomy and four cross-modal pairs including ultrasound. When combined with existing registration solvers, our plug-and-play registration measure is competitive with established end-to-end methods that design solvers and measures together.

AI USE STATEMENT

We have used AI for the following aspects of our work: For the implementation of our experiments, we partially relied on agentic coding tools. All generated code was checked manually by at least one author, and its output was sanity checked if applicable. For instance, we used coding tools to wrap baseline methods for our modular registration setup and checked that their outputs are numerically equal to running the original code. We also used coding tools to generate figures and tables and checked that the displayed information is equal to the output logs of our experiments. We used chatbots with web-search capabilities to propose related work beyond our manual literature review. We also used chatbots to suggest directions for our theoretical analysis. We drafted the proofs and had them reviewed by a chatbot. The writing of the final proofs was polished by a chatbot and verified by the authors. During the writing process, we drafted the sections in the main manuscript ourselves and employed AI agents to iterate over the drafts for some sections before writing the final versions ourselves.

Our synthetic data is generated using a procedural rendering pipeline. Therefore, AI-based data generation does not apply to our work.

We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All experiments used publicly available medical imaging datasets in accordance with their licenses and terms of use. Our method itself does not raise ethical concerns.

REPRODUCIBILITY STATEMENT

Our code and trained weights are available at `https://anonymous.4open.science/r/sigmi-3756`. Our training data is generated using publicly available work (Valabregue et al., 2026). All experimental data is publicly available or can be generated using our code with the included random seeds (Czyż et al., 2025; Dufumier et al., 2022; Taha et al., 2023; Marcus et al., 2007; Hering et al., 2022; Xiao et al., 2017; Ndzimbong et al., 2025). In addition to the code, appendix C describes experimental setups in detail. Consequently, all experiments can be reproduced up to floating point errors caused by non-deterministic code execution on GPUs.

REFERENCES

Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 924–931. Springer, 2006.

B.B. Avants, C.L. Epstein, M. Grossman, and J.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. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2007.06.004. Special Issue on The Third International Workshop on Biomedical Image Registration – WBIR 2006.

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.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pp. 531–540. PMLR, 2018.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023.

M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. Monai: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

Junyu Chen, Shuwen Wei, Joel Honkamaa, Pekka Marttinen, Hang Zhang, Min Liu, Yichao Zhou, Zuopeng Tan, Zhuoyuan Wang, Yi Wang, et al. Beyond the lumir challenge: The pathway to foundational registration models. *Medical Image Analysis*, pp. 104175, 2026.

Alexander I Cowen-Rivers, Wenlong Lyu, Rasul Tutunov, Zhi Wang, Antoine Grosnit, Ryan Rhys Griffiths, Alexandre Max Maraval, Hao Jianye, Jun Wang, Jan Peters, et al. Hebo: Pushing the limits of sample-efficient hyper-parameter optimisation. *Journal of Artificial Intelligence Research*, 74:1269–1349, 2022.

Paweł Czyż, Frederic Grabowski, Julia E Vogt, Niko Beerenwinkel, and Alexander Marx. On the properties and estimation of pointwise mutual information profiles. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=LdflD41Gn8.

Başar Demir, Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, Ebrahim Ebrahim, and Marc Niethammer. Multigradicon: A foundation model for multimodal medical image registration. In *International Workshop on Biomedical Image Registration*, pp. 3–18. Springer, 2024.

Neel Dey, Jo Schlemper, Seyed Sadegh Mohseni Salehi, Bo Zhou, Guido Gerig, and Michal Sofka. Contrareg: Contrastive learning of multi-modality unsupervised deformable image registration. In *International conference on medical image computing and computer-assisted intervention*, pp. 66–77. Springer, 2022.

Neel Dey, Mazdak Abulnaga, Benjamin Billot, Esra Abaci Turk, Ellen Grant, Adrian V Dalca, and Polina Golland. Anystar: Domain randomized universal star-convex 3d instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7593–7603, 2024.

Neel Dey, Benjamin Billot, Hallee Wong, Clinton Wang, Mengwei Ren, Ellen Grant, Adrian Dalca, and Polina Golland. Learning general-purpose biomedical volume representations using randomized synthesis. In *International Conference on Learning Representations*, volume 2025, pp. 32033–32064, 2025.

Reuben Dorent, Polina Golland, and William (Sandy) Wells. Connecting jensen–shannon and kullback–leibler divergences: A new bound for representation learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 48356–48389. Curran Associates, Inc., 2025. doi: 10.52202/085713-1614.

Benoit Dufumier, Antoine Grigis, Julie Victor, Corentin Ambroise, Vincent Frouin, and Edouard Duchesnay. Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage*, 263:119637, 2022.

Lun Gong, Cheng Zhang, Luwen Duan, Xueying Du, Hanqiu Liu, Xinjian Chen, and Jian Zheng. Nonrigid image registration using spatially region-weighted correlation ratio and gpu-acceleration. *IEEE journal of biomedical and health informatics*, 23(2):766–778, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Mattias P Heinrich, Mark Jenkinson, Manav Bhushan, Tahreema Matin, Fergus V Gleeson, Michael Brady, and Julia A Schnabel. Mind: Modality independent neighbourhood descriptor for multimodal deformable registration. *Medical image analysis*, 16(7):1423–1435, 2012.

Mattias Paul Heinrich, Mark Jenkinson, Bartlomiej W Papież, Sir Michael Brady, and Julia A Schnabel. Towards realtime multimodal fusion for image-guided interventions using self-similarities. In *International conference on medical image computing and computer-assisted intervention*, pp. 187–194. Springer, 2013.

Alessa Hering, Lasse Hansen, Tony CW Mok, Albert CS 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.

Malte Hoffmann, Benjamin Billot, Douglas N Greve, Juan Eugenio Iglesias, Bruce Fischl, and Adrian V Dalca. Synthmorph: learning contrast-invariant registration without acquired images. *IEEE transactions on medical imaging*, 41(3):543–558, 2021.

Joel Honkamaa and Pekka Marttinen. New multimodal similarity measure for image registration via modeling local functional dependence with linear combination of learned basis functions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 399–408. Springer, 2025.

Rohit Jena, Pratik Chaudhari, and James C Gee. Adaptive riemannian optimization for multi-scale diffeomorphic matching. *Nature Communications*, 17(1):4774, 2026.

Parikshit Juvekar, Reuben Dorent, Fryderyk Kögl, Erickson Torio, Colton Barr, Laura Rigolo, Colin Galvin, Nick Jowkar, Anees Kazi, Nazim Haouchine, et al. Remind: The brain resection multi-modal imaging database. *Scientific data*, 11(1):494, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Stefan Klein, Uulke A Van Der Heide, Irene M Lips, Marco Van Vulpen, Marius Staring, and Josien PW Pluim. Automatic segmentation of the prostate in 3d mr images by atlas matching using localized mutual information. *Medical physics*, 35(4):1407–1417, 2008.

Chungpa Lee, Joonhwan Chang, and Jy-yong Sohn. Analysis of using sigmoid loss for contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1747–1755. PMLR, 2024.

Nunzio A Letizia, Nicola Novello, and Andrea M Tonello. Mutual information estimation via f-divergence and data derangements. *Advances in Neural Information Processing Systems*, 37: 105114–105150, 2024.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. *Advances in neural information processing systems*, 27, 2014.

Ruizhi Liao, Daniel Moyer, Polina Golland, and William M Wells. Demi: Discriminative estimator of mutual information. *arXiv preprint arXiv:2010.01766*, 2020.

Dirk Loeckx, Pieter Slagmolen, Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Nonrigid image registration using conditional mutual information. *IEEE transactions on medical imaging*, 29(1):19–29, 2009.

Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on Medical Imaging*, 16(2):187–198, 1997.

Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. 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(9):1498–1507, 2007.

David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellyn, and William Eubank. Non-rigid multimodality image registration. In *Medical imaging 2001: image processing*, volume 4322, pp. 1609–1620. Spie, 2001.

Tony CW Mok, Zi Li, Yunhao Bai, Jianpeng Zhang, Wei Liu, Yan-Jie Zhou, Ke Yan, Dakai Jin, Yu Shi, Xiaoli Yin, et al. Modality-agnostic structural image representation learning for deformable multi-modality medical image registration. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11215–11225. IEEE, 2024.

Abhishek Nan, Matthew Tennant, Uriel Rubin, and Nilanjan Ray. Drmime: Differentiable mutual information and matrix exponential for multi-resolution image registration. In *Medical Imaging with Deep Learning*, pp. 527–543. PMLR, 2020.

William Ndzimbong, Cyril Fourniol, Loic Themyr, Nicolas Thome, Yvonne Keeza, Benoît Sauer, Pierre-Thierry Piéchaud, Arnaud Méjean, Jacques Marescaux, Daniel George, et al. Trusted: The paired 3d transabdominal ultrasound and ct human data for kidney segmentation and registration research. *Scientific Data*, 12(1):615, 2025.

XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Nicolas Pielawski, Elisabeth Wetzer, Johan Öfverstedt, Jiahao Lu, Carolina Wählby, Joakim Lindblad, and Natasa Sladoje. Comir: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33:18433–18444, 2020.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pp. 5171–5180. PMLR, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Rhodes, Kai Xu, and Michael Gutmann. Telescoping density-ratio estimation. *Advances in neural information processing systems*, 33:4905–4916, 2020.

Hassan Rivaz, Zahra Karimaghaloo, Vladimir S Fonov, and D Louis Collins. Nonrigid registration of ultrasound and mri using contextual conditioned mutual information. *IEEE transactions on medical imaging*, 33(3):708–725, 2013.

Alexis Roche, Grégoire Malandain, Xavier Pennec, and Nicholas Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 1115–1124. Springer, 1998.

Matteo Ronchetti, Wolfgang Wein, Nassir Navab, Oliver Zettinig, and Raphael Prevost. Disa: Differentiable similarity approximation for universal multimodal registration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 761–770. Springer, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Daniel B Russakoff, Carlo Tomasi, Torsten Rohlfing, and Calvin R Maurer Jr. Image similarity using mutual information of regions. In *European conference on computer vision*, pp. 596–607. Springer, 2004.

Hanna Siebert, Christoph Großbröhmer, Lasse Hansen, and Mattias P Heinrich. Convexadam: Self-configuring dual-optimization-based 3d multitask medical image registration. *IEEE Transactions on Medical Imaging*, 44(2):738–748, 2024.

Martin Simonovsky, Benjamín Gutiérrez-Becker, Diana Mateus, Nassir Navab, and Nikos Komodakis. A deep metric for multimodal registration. In *International conference on medical image computing and computer-assisted intervention*, pp. 10–18. Springer, 2016.

Gerard Snaauw, Michele Sasdelli, Gabriel Maicas, Stephan Lau, Johan Verjans, Mark Jenkinson, and Gustavo Carneiro. Mutual information neural estimation for unsupervised multi-modal registration of brain images. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 3510–3513. IEEE, 2022.

Jiaming Song and Stefano Ermon. Understanding the limitations of variational mutual information estimators. *arXiv preprint arXiv:1910.06222*, 2019.

C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999. ISSN 0031-3203. doi: https://doi.org/10.1016/S0031-3203(98)00091-0.

Alaa Taha, Greydon Gilmore, Mohamad Abbass, Jason Kai, Tristan Kuehn, John Demarco, Geetika Gupta, Chris Zajner, Daniel Cao, Ryan Chevalier, et al. Magnetic resonance imaging datasets with anatomical fiducials for quality control and registration. *Scientific Data*, 10(1):449, 2023.

Yao-Hung Hubert Tsai, Han Zhao, Makoto Yamada, Louis-Philippe Morency, and Russ R Salakhutdinov. Neural methods for point-wise dependency estimation. *Advances in Neural Information Processing Systems*, 33:62–72, 2020.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Nicholas J. Tustison, Philip A. Cook, Andrew J. Holbrook, Hans J. Johnson, John Muschelli, Gabriel A. Devenyi, Jeffrey T. Duda, Sandhitsu R. Das, Nicholas C. Cullen, Daniel L. Gillen, Michael A. Yassa, James R. Stone, James C. Gee, and Brian B. Avants. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*, 11:9068, 2021. doi: 10.1038/s41598-021-87564-6.

Romain Valabregue, Ines Khemir, Eric Badinet, François Rousseau, Guillaume Auzias, and Reuben Dorent. Siam: Head and brain mri segmentation from few high-quality templates via synthetic training. *arXiv preprint arXiv:2605.02737*, 2026.

Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

Wolfgang Wein, Alexander Ladikos, Bernhard Fuerst, Amit Shah, Kanishka Sharma, and Nassir Navab. Global registration of ultrasound to mri using the lc2 metric for enabling neurosurgical guidance. In Kensaku Mori, Ichiro Sakuma, Yoshinobu Sato, Christian Barillot, and Nassir Navab (eds.), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pp. 34–41, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

Yiming Xiao, Maryse Fortin, Geirmund Unsgård, Hassan Rivaz, and Ingerid Reinertsen. Re trospective evaluation of cerebral tumors (resect): A clinical database of pre-operative mri and intra-operative ultrasound in low-grade glioma surgeries. *Medical physics*, 44(7):3875–3882, 2017.

Yiming Xiao, Hassan Rivaz, Matthieu Chabanas, Maryse Fortin, Ines Machado, Yangming Ou, Mattias P Heinrich, Julia A Schnabel, Xia Zhong, Andreas Maier, et al. Evaluation of mri to ultrasound registration methods for brain shift correction: the curious2018 challenge. *IEEE transactions on medical imaging*, 39(3):777–786, 2019.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

A PROOFS

Proposition 1 (Identification). *Let the B pairs of a batch be i.i.d. draws from p_{UV}. Then $\mathbb{E}\big[\mathcal{L}_{\mathrm{SigLIP}}\big] = B\,\mathcal{L}_\pi(s)$ with $\pi = 1/B$. More generally, weighting diagonal terms by π/B and off-diagonal terms by $(1-\pi)/\big(B(B-1)\big)$ yields $\mathcal{L}_\pi(s)$ for any $\pi \in (0,1)$ at fixed B.*

Proof. Each matched pair (u_i, v_i) is a draw from p_{UV}. For $i \neq j$, u_i and v_j belong to independent pairs, so (u_i, v_j) is a draw from $p_U \otimes p_V$. Group equation (4) into its B diagonal and $B(B-1)$ off-diagonal terms and divide by B: $\mathcal{L}_{\mathrm{SigLIP}}/B = \frac{1}{B}\big[\frac{1}{B}\sum_i -\log\sigma(s_{ii})\big] + \frac{B-1}{B}\big[\frac{1}{B(B-1)}\sum_{i\neq j} -\log(1-\sigma(s_{ij}))\big]$. The diagonal average is therefore unbiased for $\mathbb{E}_{p_{UV}}[-\log\sigma(s)]$ and the off-diagonal average for $\mathbb{E}_{p_U\otimes p_V}[-\log(1-\sigma(s))]$. Their weights are $1/B$ and $(B-1)/B$, which is equation (6) at $\pi = 1/B$. The general case follows by substituting the stated weights. □

Theorem 2 (Bayes-optimal score). *For every $\pi \in (0,1)$, the minimizer of $\mathcal{L}_\pi$ over measurable scores is, m_π-almost everywhere,*

$$s^\star_\pi(u,v) \;=\; \mathrm{PMI}(u,v) \;+\; \log\frac{\pi}{1-\pi}. \tag{7}$$

Proof. Conditioning equation (6) on (U,V) and writing $\eta = \mathrm{d}(\pi p_{UV})/\mathrm{d}m_\pi = \mathbb{P}(Z{=}1 \mid U,V)$,

$$\mathcal{L}_\pi(s) = \mathop{\mathbb{E}}_{m_\pi}\Big[-\eta\log\sigma(s) - (1-\eta)\log\big(1-\sigma(s)\big)\Big], \tag{14}$$

whose integrand is, for each fixed (u,v), the cross-entropy between $\mathcal{B}(\eta)$ and $\mathcal{B}(\sigma(s))$. By Gibbs' inequality it is minimized uniquely at $\sigma(s) = \eta$, and the minimization is pointwise, so $\sigma(s^\star_\pi) = \eta$ m_π-a.e. Applying the logit and Bayes' rule, $s^\star_\pi = \log\frac{\eta}{1-\eta} = \log\frac{\pi\,\mathrm{d}p_{UV}}{(1-\pi)\,\mathrm{d}p_U\otimes p_V} = \mathrm{PMI} + \log\frac{\pi}{1-\pi}$. □

Theorem 5 (Excess risk is a Bregman divergence in the logit). *With $A(z) = \log(1+e^z)$,*

$$\Delta_\pi(s) \;=\; \mathop{\mathbb{E}}_{m_\pi}\big[A\big(s^\star_\pi+\varepsilon\big) - A\big(s^\star_\pi\big) - \varepsilon\,\sigma\big(s^\star_\pi\big)\big] \;=\; \tfrac{1}{2}\mathop{\mathbb{E}}_{m_\pi}\big[\sigma'(\xi)\,\varepsilon^2\big] \tag{11}$$

for some ξ between $s^\star_\pi$ and s. If moreover $|s|, |s^\star_\pi| \le M$, then with $c_M = \sigma(M)\big(1-\sigma(M)\big)$,

$$\tfrac{c_M}{2}\mathop{\mathbb{E}}_{m_\pi}\big[\varepsilon^2\big] \;\le\; \Delta_\pi(s) \;\le\; \tfrac{1}{8}\mathop{\mathbb{E}}_{m_\pi}\big[\varepsilon^2\big]. \tag{12}$$

Proof. By equation (14), $\Delta_\pi(s) = \mathbb{E}_{m_\pi}\big[D_{\mathrm{KL}}(\mathcal{B}(\eta)\,\|\,\mathcal{B}(\sigma(s)))\big]$, since subtracting the risk at $\sigma(s) = \eta$ replaces each cross-entropy by the corresponding KL divergence. Parameterizing the Bernoulli family by its natural parameter, with log-partition A and mean $A' = \sigma$, the KL divergence between members with logits θ_1, θ_2 is the Bregman divergence $A(\theta_2) - A(\theta_1) - (\theta_2-\theta_1)A'(\theta_1)$; taking $\theta_1 = s^\star_\pi$, $\theta_2 = s$ gives the first equality. Taylor's theorem with Lagrange remainder and $A'' = \sigma'$ gives the second. For equation (12), $\sigma'(\xi) = \sigma(\xi)(1-\sigma(\xi)) \le 1/4$ always, and $\sigma'(\xi) \ge c_M$ whenever $\xi \in [-M, M]$, which holds because ξ lies between $s^\star_\pi$ and s. □

Corollary 6 (Error under the joint, and the price of imbalance). *Under the hypotheses of theorem 5,*

$$\mathop{\mathbb{E}}_{p_{UV}}\Big[\big(\widehat{\mathrm{PMI}} - \mathrm{PMI}\big)^2\Big] \;\le\; \frac{1}{\pi}\mathop{\mathbb{E}}_{m_\pi}\big[\varepsilon^2\big] \;\le\; \frac{2\,\Delta_\pi(s)}{c_M\,\pi}. \tag{13}$$

Proof. $m_\pi \ge \pi p_{UV}$, hence $\mathrm{d}p_{UV}/\mathrm{d}m_\pi \le 1/\pi$ and $\mathbb{E}_{p_{UV}}[\varepsilon^2] \le \pi^{-1}\mathbb{E}_{m_\pi}[\varepsilon^2]$; apply the left inequality of equation (12). □

B IMPLEMENTATION DETAILS

B.1 SYNTHETIC TRAINING DATA

Synthetic dataset. We aim at an estimator that handles arbitrary medical image modalities. A recent line of work shows models trained on synthetic, domain-randomized data transfer well to real modalities (Hoffmann et al., 2021; Billot et al., 2023; Dey et al., 2024; 2025; Valabregue et al., 2026). The core idea is to generate samples by starting with anatomical segmentation maps or fully randomized templates, assigning random intensity ranges per segmentation class and per image, and applying strong appearance augmentations. For discriminative approaches, perfectly aligned image pairs are created by sharing the same segmentation template but assigning different contrast models per image. We follow Valabregue et al. (2026), generating all training data from just five brain segmentation maps. We generate 1,337 paired volumes in total and split them into 1,134 training pairs and 203 hold-out pairs for testing during development. We orient all volumes to a common coordinate system and scale their intensities to $[0, 1]$.

Batch construction. Although our fully convolutional architecture supports whole images, we train it on local patches that match each variant's receptive field size to increase the data diversity per batch. Each batch is constructed from multiple image pairs, each contributing multiple local patch pairs. For each image pair, we independently sample one isotropic voxel spacing uniformly between 1.5 and 3.0 mm . We choose this range to ensure that patches still contain heterogeneous intensities at the lower end while remaining visually abstract to prevent overfitting brain structure at the higher end . Within each image pair, patches are drawn from a randomly offset grid to guarantee all positive patch pairs are perfectly aligned and all formed negatives share zero spatial overlap. We sample the patches as spatially connected groups from the grid to promote sampling pairs from visually similar areas, resulting in harder negatives. All crops lie fully inside the resampled volumes. Since the number of local patches constructable from one image pair varies with image size, voxel spacing, and grid placement, we construct our batches by drawing a variable number of image pairs without replacement and limit the maximum number of patches drawn from one image to 16.

Augmentation. We apply the same spatial transformations to both elements of a patch pair, preserving their correspondence. Axis flips, a random right-angle rotation, and an affine transformation with rotations up to $\pm 30^\circ$, shear coefficients up to ± 0.15, translations up to ± 5 voxels, and scale factors between 0.75 and 1.25 along each axis are each applied independently with probability 0.5. The affine transformation and crop are evaluated lazily together, so the warp can use surrounding volume content and does not generally create zero borders. Only locations outside the volume are filled with zeros. These augmentations can occasionally introduce small overlapping regions between patch pairs or produce effective voxel spacings outside our defined range, which we deliberately accept as minor training noise to simplify our data pipeline . Finally, we independently augment the patches of each pair using zero-mean Gaussian noise with $\sigma \sim \mathcal{U}(0, 0.1)$, Gaussian smoothing with a per-axis $\sigma \sim \mathcal{U}(0.25, 1.5)$ voxels, and a third-degree smooth bias field with coefficients sampled from $\mathcal{U}(-0.1, 0.1)$, each with probability 0.5.

Field-of-view masking. In density-ratio estimation, trivially separable joint and product of marginals distributions can result in strong classifiers that still yield poor estimates (Rhodes et al., 2020). Especially at higher voxel spacings, many of our training patches fall on the border of anatomical structure and background. In those cases, the foreground silhouette can act as a shortcut for identifying correspondence, making classification trivial. To break this shortcut, with probability 0.7, we apply independent field-of-view masks to patches from a pair. Each mask retains only the interior of a randomly oriented, smoothly deformed ellipsoid. We sample masks in normalized patch coordinates $x \in [-1, 1]^3$. The rotation $Q = [n, t_1, t_2]$, sampled uniformly from $\mathrm{SO}(3)$, gives the ellipsoid's principal-axis frame: n is its normal direction and t_1, t_2 span its tangent plane. The radius r_n controls the ellipsoid extent along n, while $r_{t,1}$ and $r_{t,2}$ control its extent along the tangent axes. We sample these radii using the bounded power-uniform distribution $\mathcal{P}(\ell, h, \gamma) = \ell + (h - \ell)U^\gamma$ for $U \sim \mathcal{U}(0, 1)$. The radii are sampled as

$$r_n \sim \mathcal{P}(r_n^{\min}, r_n^{\max}, \gamma_r), \qquad r_{t,1}, r_{t,2} \overset{\text{iid}}{\sim} \mathcal{P}(r_t^{\min}, r_t^{\max}, \gamma_r). \tag{15}$$

To couple ellipsoid size to where its surface crosses the patch, we first summarize its size by the normalized mean radius $\bar{r} \in [0, 1]$:

$$\bar{r} \triangleq \frac{1}{3}\left(\frac{r_n - r_n^{\min}}{r_n^{\max} - r_n^{\min}} + \sum_{j=1}^{2} \frac{r_{t,j} - r_t^{\min}}{r_t^{\max} - r_t^{\min}}\right). \tag{16}$$

We then construct the anchor a, the point inside the patch at which the ellipsoid surface will intersect the normal. The vector $n/\|n\|_\infty$ reaches the patch boundary along the normal direction, and the anchor fraction λ sets how far along this ray a lies. We construct λ to be positively coupled with the ellipsoid size. This places the low-curvature surface of large ellipsoids near the patch boundary, where it resembles a smooth cutting plane, and the surface of small ellipsoids closer to the patch center to produce rounder regions that can intersect several patch faces. Specifically,

$$\begin{aligned} \lambda \triangleq \lambda_{\min} + (\lambda_{\max} - \lambda_{\min})\left(\kappa\bar{r} + (1-\kappa)U_a^{\gamma_a}\right), \qquad U_a \sim \mathcal{U}(0,1), \\ a \triangleq \lambda \frac{n}{\|n\|_\infty}. \end{aligned} \tag{17}$$

To simulate realistic silhouettes, we apply strong non-linear deformations to the ellipsoid. The vector $\psi(x)$ contains quadratic and cubic monomials of the normalized coordinate x and provides a smooth spatial basis for our deformation:

$$\begin{aligned} \psi(x) \triangleq (x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3, x_1^3, x_2^3, x_3^3, \\ x_1^2x_2, x_1^2x_3, x_2^2x_1, x_2^2x_3, x_3^2x_1, x_3^2x_2, x_1x_2x_3)^\top. \end{aligned} \tag{18}$$

We sample smooth displacement fields $d(x)$ of varying strengths as:

$$A \sim \mathcal{P}(A_{\min}, A_{\max}, \gamma_d), \qquad G_{kj} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad d(x) \triangleq A\frac{G^\top\psi(x)}{\|G\|_{\text{F}}}. \tag{19}$$

We define the ellipsoid center c such that the deformed surface passes through the anchor a:

$$c \triangleq a + d(a) - r_n n. \tag{20}$$

Finally, we apply our mask by retaining all voxels contained in the ellipsoid:

$$M(x) \triangleq \mathbf{1}\left[\sum_{j=1}^{3}\left(\frac{[Q^\top(x + d(x) - c)]_j}{r_j}\right)^2 \le 1\right]. \tag{21}$$

We heuristically tune masking parameters to produce smooth, moderately deformed cutting planes and corners and set $(r_n^{\min}, r_n^{\max}) = (1.25, 8)$, $(r_t^{\min}, r_t^{\max}) = (0.75, 8)$, $\gamma_r = 0.7$, $(\lambda_{\min}, \lambda_{\max}) = (0.2, 0.95)$, $\kappa = 0.65$, $\gamma_a = 0.6$, $(A_{\min}, A_{\max}) = (0.15, 3)$, and $\gamma_d = 1.5$. We reject and resample mask pairs whose intersection retains less than 25% of the patch voxels.

B.2 Shared, fully convolutional 3D encoder.

Encoder architecture. We use fully convolutional 3D encoders in the form of a residual backbone (He et al., 2016) with four stages with channel widths (d_1, d_2, d_3, d_4) consisting of two residual blocks each, and a final embedding projection head. Each backbone block has two convolutions: the first followed by a ReLU, and the second followed by the residual addition and a ReLU, totaling 8 blocks and 16 convolutions and ReLU applications for the backbone. Convolutions use unit stride, unit dilation, no padding, and no bias. Each skip connection is center-cropped to match the valid convolutions and uses a bias-free pointwise projection when its channel width changes. We use no pooling, spatial downsampling, or spatial normalization. A pointwise embedding head with bias maps the backbone output to the final embeddings of size d_{emb}, which we L_2-normalize. All residual projections and the final embedding head are realized using $1 \times 1 \times 1$ convolutions.

Let R denote the side length of the encoder's cubic receptive field. Each encoder accepts inputs with spatial dimensions $(x, y, z) \in \mathbb{N}^3_{\ge R}$ and produces a dense embedding field of shape $(x - R + 1, y - R + 1, z - R + 1, d_{\text{emb}})$, with one embedding computed from each valid $R \times R \times R$ input window. For isotropic convolution kernel sizes $(k_{b,1}, k_{b,2})$ in block b, $R = 1 + \sum_{b=1}^{8}[(k_{b,1} - 1) + (k_{b,2} - 1)]$. We apply the same kernel sizes $(k_{1,1}, k_{1,2})$; $(k_{2,1}, k_{2,2})$ to each backbone stage to construct encoder variants with target Rs. Appendix table 2 shows our SigMI variants.

Table 2: **SigMI variants for medical image registration.** Kernel sizes give the side lengths of the two convolutions in the first and second block of each stage. Parameter counts include the final embedding projection and the learned scale t and bias b.

	SigMI-R9	SigMI-R17
Receptive field R	9	17
Kernel sizes	$(3,1);(1,1)$	$(3,1);(3,1)$
Stage widths	$(16,32,64,128)$	$(16,32,64,128)$
Parameters	383,298	949,058
Scale t (nats)	8.459	12.777
Bias b (nats)	-5.780	-6.653
Class prior π	$1/32$	$1/32$
Prior correction (nats)	3.434	3.434
PMI range (nats)	$[-10.806,\ 6.113]$	$[-15.996,\ 9.558]$

Image registration loss. For image registration, we first compute valid embedding fields for both images. At test time, we parameterize the last encoder block's second convolution's stride s, which is equal to computing the dense embedding grid ($s = 1$) and subsampling to every sth embedding along all dimensions. Setting higher s reduces computation and memory requirements at the cost of slightly worse similarity estimates. We use equation (9) to obtain a field of PMI estimates, whose negative mean defines our image-wise registration loss.

Let I and J denote the two images, and let m_I and m_J be their binary region-of-interest masks with matching spatial dimensions. When no masks are provided, both masks are identically one. Let $\Omega^{(s)}$ be the set of valid voxel centers retained at stride s. Writing $N_{I,J}^{(s)} \triangleq \sum_{\mathbf{x}\in\Omega^{(s)}} m_I(\mathbf{x})m_J(\mathbf{x})$ for the number of jointly valid centers, we define

$$\mathcal{L}_{\mathrm{SigMI}}(I,J,m_I,m_J;s) \;=\; -\frac{1}{N_{I,J}^{(s)}} \sum_{\mathbf{x}\in\Omega^{(s)}} m_I(\mathbf{x})\, m_J(\mathbf{x})\, \widehat{\mathrm{PMI}}_{I,J}(\mathbf{x}), \tag{22}$$

where $\widehat{\mathrm{PMI}}_{I,J}(\mathbf{x})$ is the local PMI estimate at center $\mathbf{x}$.

Training. We train with batches of $B = 32$ corresponding patch pairs and use all within-batch mismatches, giving $\pi = 1/32$. We accumulate gradients over four independent batches before each parameter update. We initialize convolutional weights with Kaiming-normal initialization. The scale is initialized to $t = 10$ and optimized in log space to keep it positive. The model bias is initialized to $b = -\log(B-1) = -\log 31$. We jointly optimize the encoder, scale, and bias for 100,000 steps using Adam with an initial learning rate of 10^{-4} and weight decay 10^{-5} applied to all parameters except t and b . We decay the learning rate to 10^{-6} with a cosine schedule and use the final checkpoint for inference.

C Experiment details

C.1 PMI estimation

Estimators. We perform our experiments on the BMI v2 (Czyż et al., 2025) Bend and Mix Model tasks with the following estimators: *NWJ* (Nguyen et al., 2010) is a neural estimator trained to maximize $\mathbb{E}_{p_{UV}}[f_\theta(u,v)] - \mathbb{E}_{p_U p_V}[\exp(f_\theta(u,v)-1)]$, applying an exponential penalty to samples of the product of marginals. At test time, assuming an optimal critic, pairwise PMI estimates can be obtained as $\widehat{\mathrm{PMI}}_{\mathrm{NWJ}}(u,v) = f_\theta(u,v) - 1$. Mutual information estimates are obtained by correcting the critic's normalization: $\widehat{I}(U;V) \;=\; \frac{1}{N}\sum_{i=1}^{N} f_\theta(u_i,v_i) - \frac{1}{N(N-1)}\sum_{i\neq j}\exp\big(f_\theta(u_i,v_j)-1\big)$. We implement NWJ following BMI's baseline implementation and training recipe. *InfoNCE* (Oord et al., 2018) learns a critic by maximizing the joint partner's score relative to other marginal samples. At test time, PMI estimates can be computed by normalizing the critic's score with respect to scores it assigns to unpaired samples constructed using a reference set $\mathcal{R}$: $\widehat{\mathrm{PMI}}_{\mathrm{NCE}}(u_i,v_i) = f_\theta(u_i,v_i) -$

Table 3: **Geometric BMI v2 tasks.** Entries show mean (standard deviation) across ten experiment seeds at $B = 256$. Bold means indicate the closest MI estimate to ground truth or the lowest sample-wise PMI L_1 error. GT MI denotes ground-truth mutual information, MI its estimate, and L_1 the sample-wise PMI estimation error. All values are in nats.

Metric	Estimator	X	AI	Waves	Galaxy
GT MI		0.409	0.778	1.308	0.493
MI	NWJ	0.372 (0.012)	0.512 (0.095)	0.132 (0.077)	0.262 (0.068)
	InfoNCE	**0.399** (0.013)	0.728 (0.020)	**0.753** (0.104)	**0.418** (0.016)
	SigLIP	−5.154 (0.021)	−4.812 (0.055)	−4.856 (0.098)	−5.131 (0.031)
	SigMI	0.387 (0.021)	**0.729** (0.055)	0.686 (0.098)	0.411 (0.031)
L_1	NWJ	0.174 (0.021)	0.474 (0.096)	1.234 (0.089)	0.389 (0.075)
	InfoNCE	**0.089** (0.011)	**0.197** (0.014)	**0.654** (0.092)	**0.212** (0.018)
	SigLIP	5.564 (0.024)	5.590 (0.047)	6.160 (0.101)	5.625 (0.031)
	SigMI	0.106 (0.011)	0.209 (0.014)	0.740 (0.075)	0.234 (0.021)

Table 4: **Concentric BMI v2 tasks.** Entries show mean (standard deviation) across ten experiment seeds at $B = 256$. Bold means indicate the closest MI estimate to ground truth or the lowest sample-wise PMI L_1 error. GT MI denotes ground-truth mutual information, MI its estimate, and L_1 the sample-wise PMI estimation error. All values are in nats.

		3D		5D		25D
Metric	Estimator	5 comp.	10 comp.	5 comp.	10 comp.	5 comp.
GT MI		0.456	0.556	0.626	0.749	1.190
MI	NWJ	0.376 (0.021)	0.447 (0.024)	0.475 (0.084)	0.552 (0.082)	0.506 (0.192)
	InfoNCE	**0.420** (0.011)	**0.516** (0.015)	0.571 (0.017)	**0.685** (0.014)	0.572 (0.037)
	SigLIP	−5.132 (0.039)	−5.030 (0.032)	−4.967 (0.073)	−4.860 (0.053)	−4.939 (0.076)
	SigMI	0.409 (0.039)	0.511 (0.032)	**0.574** (0.073)	0.681 (0.053)	**0.603** (0.076)
L_1	NWJ	0.282 (0.042)	0.331 (0.041)	0.378 (0.094)	0.426 (0.114)	0.893 (0.176)
	InfoNCE	**0.168** (0.026)	**0.177** (0.011)	**0.213** (0.026)	**0.239** (0.026)	0.791 (0.034)
	SigLIP	5.585 (0.038)	5.586 (0.033)	5.593 (0.070)	5.612 (0.049)	6.127 (0.076)
	SigMI	0.168 (0.020)	0.185 (0.012)	0.227 (0.028)	0.242 (0.019)	**0.764** (0.051)

$\log\left[\frac{1}{|\mathcal{R}|}\sum_{v'\in\mathcal{R}}\exp f_\theta(u_i, v')\right]$. MI estimates are obtained by taking their expectation over the joint. We implement the InfoNCE critic using separate MLP encoders for U and V. Each encoder has two hidden layers with 8 and 32 channels and ReLU activations, followed by a projection to an 8-dimensional representation. This architecture is fixed across tasks. We normalize the embeddings to unit length and compute $f_\theta(u, v)$ as their dot product multiplied by a learned positive temperature, initialized to 10. We use a learning rate of 0.01 and an early stopping patience of 8 checks, but retain BMI's remaining training parameters and InfoNCE objective. *SigLIP* uses the same encoder-based architecture as InfoNCE, but computes its score with a learned bias as shown in equation (3) and is trained with sigmoid loss. It is not a calibrated PMI estimator and its scores are offset by $\log(B-1)$. *SigMI* is our estimator that yields calibrated PMI and MI estimates. It uses the same trained encoder and affine weights as SigLIP, but computes scores using the bias correction shown in equation (8).

Bend and Mix Model tasks. We follow BMI's released setup and draw 5,000 joint pairs for each task, which we split into 2,500 training and 2,500 test samples. In addition to BMI's reported ground-truth MI and estimated MI, we report L_1 errors for PMI estimates. All experiments are repeated with 10 random seeds and we report their mean metrics. Results for the 13 Bend and Mix Model tasks are shown in appendix tables 3 to 5.

Table 5: **Inlier BMI v2 tasks.** Entries show mean (standard deviation) across ten experiment seeds at $B = 256$. Bold means indicate the closest MI estimate to ground truth or the lowest sample-wise PMI L_1 error. GT MI denotes ground-truth mutual information, MI its estimate, and L_1 the sample-wise PMI estimation error. All values are in nats.

		5D		25D	
Metric	Estimator	20%	50%	20%	50%
GT MI		0.633	0.268	0.634	0.271
MI	NWJ	**0.577** (0.025)	**0.215** (0.024)	0.339 (0.182)	0.013 (0.066)
	InfoNCE	0.518 (0.024)	0.150 (0.020)	0.309 (0.058)	−0.115 (0.050)
	SigLIP	−5.035 (0.041)	−5.401 (0.036)	−5.181 (0.060)	−5.519 (0.028)
	SigMI	0.506 (0.041)	0.140 (0.036)	**0.360** (0.060)	**0.022** (0.028)
L_1	NWJ	**0.274** (0.010)	**0.251** (0.026)	0.650 (0.228)	**0.610** (0.094)
	InfoNCE	0.354 (0.030)	0.380 (0.031)	0.675 (0.046)	0.766 (0.052)
	SigLIP	5.674 (0.028)	5.671 (0.029)	5.822 (0.045)	5.795 (0.027)
	SigMI	0.365 (0.023)	0.382 (0.019)	**0.611** (0.044)	0.611 (0.038)

Reference set dependence. Compared to InfoNCE, SigMIs have the favorable property that their estimates are independent of any reference set. To illustrate this, we draw an independent reference pool of 10,000 samples and report InfoNCE's estimates with respect to reference pool subsets of varying sizes. Each sample of the test set is scored by normalizing with 10 random reference subsets of sizes $|\mathcal{R}| \in \{7, 15, 31, 63, 127, 255, 511, 1023, 2047, 4095\}$. Appendix figure 6 shows six samples from three representative tasks.

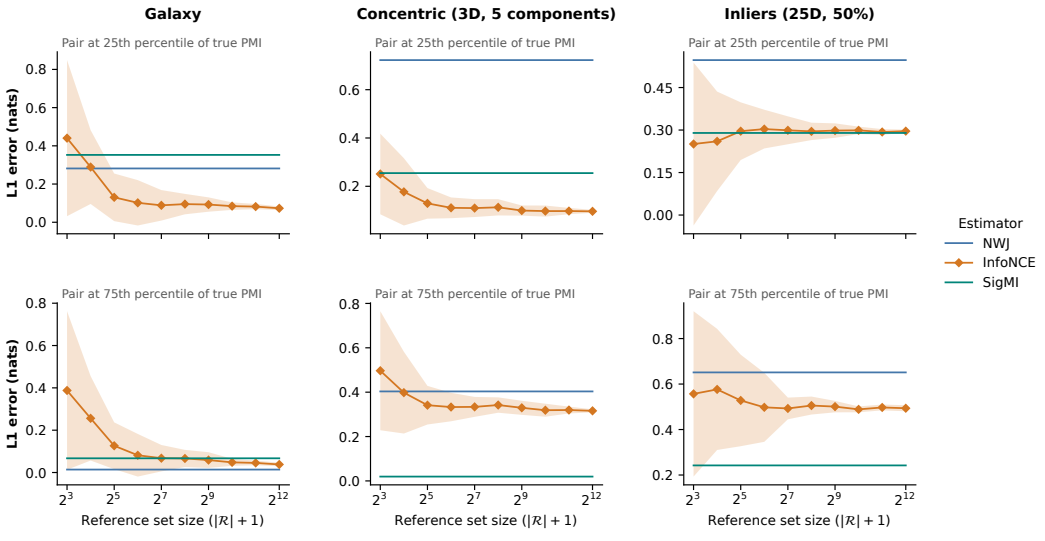

Figure 6: **Reference set dependence of pairwise PMI L_1 error.** For each task, we show fixed evaluation pairs at the 25th and 75th percentiles of ground-truth PMI. Curves indicate mean L_1 error and its standard deviation over 10 reference sets. InfoNCE's estimates show strong variance for small reference sets, while SigMI and NWJ give one reference-independent estimate per pair.

C.2 Registration datasets

Learn2Reg 2025 *LuMIR* (Chen et al., 2026; Dufumier et al., 2022; Taha et al., 2023; Marcus et al., 2007) is a brain MRI dataset that consists of 27 locally evaluable pairs formed from 38 distinct volumes, split into 9 T1 to T1 pairs, 9 high-field T1 to high-field T1 pairs, and 9 T1 to T2 pairs.

We split each category into 3 validation and 6 test samples, totaling 9 validation and 18 test set samples. One underlying volume can be part of 2 pairs. We pick our splits to minimize the number of volumes appearing in both validation and test sets, which results in one volume shared each for the T1 categories, and zero volumes shared for the multimodal category. All volumes have a voxel size of $[1.0, 1.0, 1.0]$ mm and dimensions $[160, 224, 192]$. The data is skull-stripped and underwent affine alignment. LuMIR does not provide region-of-interest masks. Its public annotations comprise segmentation maps of 10 anatomical classes. We follow the challenge's evaluation protocol and report mean Dice score.

Learn2Reg 2021 *Abdomen* (Hering et al., 2022) consists of 8 publicly available T1 MRI to CT pairs. We split them into a validation set of 4 pairs and a test set of 4 pairs. All volumes have voxel sizes of $[2.0, 2.0, 2.0]$ mm and dimensions $[192, 160, 192]$. The volumes underwent group-wise affine alignment. We perform unmasked registration with no dataset-specific preprocessing. The dataset provides segmentation maps for liver, spleen, and left and right kidneys. We follow the official evaluation protocol and report mean Dice score.

The *RESECT* brain dataset (Xiao et al., 2017) offers FLAIR MRI to intra-operative, pre-resection US pairs from 22 low-grade-glioma patients. We use 8 pairs for the validation set and 14 pairs for the test set. Native voxel spacing and volume dimensions vary across subjects and modalities: FLAIR MRI has mean voxel spacing $[1.000, 1.002, 1.002]$ mm and mean dimensions $[192.7, 251.6, 253.1]$, whereas US has mean voxel spacing $[0.214, 0.214, 0.214]$ mm and mean dimensions $[336.7, 365.0, 336.1]$ voxels. The pairs come with rigid alignment from a neuro-navigation system. We resample the ultrasound to the MRI image grid and perform masked registration with ultrasound masks generated by thresholding the ultrasound and performing $6 \times 6 \times 6$ average pooling. RESECT provides 15-16 expert-annotated landmarks per pair, 338 in total. We follow the CuRIOUS 2018 challenge's evaluation protocol (Xiao et al., 2019) and report mean TRE in millimeters.

TRUSTED (Ndzimbong et al., 2025) is a public kidney registration benchmark consisting of 59 CT and ultrasound pairs collected from 39 patients. We use its first provided fold to keep patients split between test and validation set and invert it, resulting in 12 validation and 47 test samples. Native voxel spacing and volume dimensions vary across subjects and modalities: CT has mean voxel spacing $[0.822, 0.822, 1.049]$ mm and mean dimensions $[512.0, 512.0, 599.8]$, whereas US has mean voxel spacing $[0.300, 0.300, 0.300]$ mm and mean dimensions $[1024.0, 768.0, 739.7]$ voxels. We apply the provided initial similarity transform estimated from landmarks 3 to 5, without added landmark noise. We remove embedded US preview frames and graphics, then resample CT and US onto a shared grid with voxel spacing $[2.0, 2.0, 2.0]$ mm, yielding mean dimensions of $[211.4, 211.4, 294.8]$ voxels. Similar to RESECT, we perform masked registration with US masks derived from nonzero intensities by $6 \times 6 \times 6$ average pooling. TRUSTED provides kidney segmentations and 7 expert-annotated landmarks per kidney, although Ndzimbong et al. (2025) state there is only a single reliably identifiable landmark. We therefore use a segmentation-based Dice score as our metric, which we compute based on voxels.

C.3 Registration solvers

MultiresSVF is LoCoR's (Honkamaa & Marttinen, 2025) implementation of a multi-resolution, diffeomorphic registration pipeline based on scaling and squaring (Arsigny et al., 2006). Its gradient-based optimization works with all differentiable image similarities. We fix its resolution schedule and regularizer type per dataset and search parameters regarding optimization and regularization for each dataset-similarity combination. For LuMIR and Abdomen datasets we use Honkamaa & Marttinen's (2025) released default resolution schedule and bending-energy regularizer. For RESECT and TRUSTED, we use the resolution schedule of their ReMIND2Reg 2025 submission with diffusion regularization. Since TRUSTED comes with reliable rigid and isotropically scaled alignment, we disable the included initial rigid stage in this case. For all dataset-similarity combinations, if applicable, we search Adam's learning rate for the initial affine or rigid stage within $[10^{-3}, 10^{-1}]$ in log space, the learning rate for all deformable stages within $[10^{-3}, 10^{-1}]$ in log space, and a multiplier for each stage's regularization weight in $[10^{-1}, 10^{4}]$ in log space.

ConvexAdam (Siebert et al., 2024) creates an initial deformation field based on an offset-cost volume and convex optimization, followed by further gradient-based optimization. The cost volume assigns matching costs for a set of candidate offsets at regular spatial locations. Therefore, similar-

ity measures must be locally applicable and differentiable. In the convex stage, we construct offset costs for image-based similarity measures (Squared LNCC, Mattes MI, NMI, SRWCR, SigMI) by defining a cubic, local window centered on each voxel, tracing which original resolution image voxels influence each candidate cost in ConvexAdam's original representation-based implementation, and computing a scalar matching cost for the union of all local windows centered on these voxels. We found this approach to yield the best results for heuristic experiments across similarities and datasets during development. For ConvexAdam, we fix the convex stage's spatial parameters and its smoothing schedule per dataset and search the Adam stage's learning rate and regularization weight for each dataset-similarity combination. We set the fixed parameters for LuMIR equal to the challenge's official ConvexAdam baseline (Chen et al., 2026). For Abdomen, we follow Siebert et al.'s (2024) parameters for the same dataset but approximate their Gaussian deformation smoothing with two passes of the released implementation's $3 \times 3 \times 3$ average pooling. For RESECT and TRUSTED, we use their convex stage parameters for the CuRIOUS challenge, whose publicly released data is the RESECT dataset. Inspired by Juvekar et al.'s (2024) ReMIND2Reg 2025 ConvexAdam baseline, we disable the Adam stage for both datasets and additionally constrain the resulting deformation field to an affine approximation for TRUSTED. If applicable, the Adam stage's regularization weight is searched in $[10^{-2}, 10^{2}]$ in log space and the Adam stage's learning rate in $[10^{-2}, 10^{2}]$ in log space.

FireANTs SyN (Jena et al., 2026) is a multi-resolution, diffeomorphic algorithm based on adaptive Riemannian optimization. Similar to the widely used ANTs SyN (Avants et al., 2008; Tustison et al., 2021) algorithm, it computes image similarities in a midpoint space and optimizes transformations for both images. Its Python-based implementation allows for arbitrary differentiable image similarities. We fix the resolution schedule per dataset and perform hyperparameter optimization for its Adam learning rate and the standard deviations for Gaussian gradient and warp smoothing per dataset-similarity combination. We heuristically tune the resolution schedule for all datasets, search the Adam learning rate in $[10^{-2}, 10^{1}]$ in log space, the gradient smoothing standard deviation in $[0, 5]$, and the warp smoothing standard deviation in $[0, 2.5]$. Because FireANTs SyN optimizes two transformations, its memory requirements are high. For the RESECT and TRUSTED datasets, we set stride $s = 2$ for SigMIs.

We perform all hyperparameter optimization over the validation sets using the HEBO Bayesian optimizer (Cowen-Rivers et al., 2022) with a budget of 20 trials per dataset-algorithm-similarity combination.

C.4 Similarity measures

Squared LNCC (Avants et al., 2008) computes the mean squared, normalized cross-correlation over local windows. It gives high similarity for negative correlations, allowing it to handle multimodal settings. For gradient-based registration, we use MONAI's (Cardoso et al., 2022) implementation with default parameters. To construct ConvexAdam's cost volume, we set the local window size to 17 voxels.

Mattes MI (Mattes et al., 2001) is a histogram-based implementation of mutual information that replaces Viola & Wells III's (1997) Gaussian Parzen kernels with B-splines. We use the MONAI implementation with default parameters. To compute the ConvexAdam cost volume efficiently, we implement a custom CUDA kernel that processes local windows in parallel. We pick a local window size of 17.

NMI (Studholme et al., 1999) is a variation of mutual information that normalizes with respect to the joint entropy. That makes it less sensitive to the shared spatial overlap of two images. As for Mattes MI, we use MONAI with default parameters for the gradient-based settings and our custom CUDA kernel for parallel cost volume construction. The cost volume uses window size 17.

SRWCR (Gong et al., 2018) implements a local correlation ratio using a spatially weighted histogram, causing voxels closer to the window center to contribute more. We use Honkamaa & Marttinen's (2025) baseline implementation with default parameters for gradient-based registration, and a custom implementation for ConvexAdam cost volume construction. We set its cost volume window size to 16.

MIND-SSC (Heinrich et al., 2013) is a successor of the widely used MIND (Heinrich et al., 2012) self-similarity descriptor. It replaces the intensity image with voxel-wise descriptors, whose mean squared errors then form a strong similarity measure. We use Siebert et al.'s (2024) implementation with default parameters.

Anatomix (Dey et al., 2025) learns modality-independent representations using a U-Net (Ronneberger et al., 2015) with a multi-positive supervised contrastive objective (Khosla et al., 2020) using domain-randomized, synthetic data. We use the released weights and follow its released implementation, omitting the concatenation with MIND-SSC features for the ConvexAdam cost volume construction to isolate its performance.

C.5 Full registration methods

multiGradICON (Demir et al., 2024) is a foundation registration model trained on multiple datasets spanning multiple anatomies. We apply its instance-wise finetuning with default parameters and tune with respect to its lncc^2 similarity because squared LNCC performed the best overall of all similarities implemented by multiGradICON in our LuMIR and RESECT experiments in section 4.2. Since the RESECT ultrasound images' spatial support differs drastically from MRI, we adapt multiGradICON's finetuning protocol to evaluate similarities only within the intersection of fixed and moving image masks. Even though multiGradICON was only trained on MRI and CT modalities, we report its performance on RESECT for completeness.

ANTs SyN (Tustison et al., 2021; Avants et al., 2008) is a widely used registration approach based on symmetric image normalization. It optimizes transformations for both images, computing similarities on midpoint images. We use its Mattes MI (Mattes et al., 2001) implementation. Similar to LoCoR's registration algorithm, it requires configuring stages and setting registration hyperparameters. For LuMIR, we rely on Honkamaa & Marttinen's (2025) ANTs baseline stage configuration, for RESECT we use the same stage settings as for our FireANTs SyN solver in section 4.2.

SynthMorph (Hoffmann et al., 2021) is a variant of VoxelMorph (Balakrishnan et al., 2019) trained on synthetic data consisting of domain-randomized random shapes. Because SynthMorph does not support masked registration, we don't supply RESECT's masks.

LoCoR (Honkamaa & Marttinen, 2025), similarly to SRWCR (Gong et al., 2018), is a multimodal similarity measure that approximates a local correlation ratio. It is packaged with the multiresolution, diffeomorphic registration pipeline we employed for our measure comparison in section 4.2 (MultiresSVF), and optimizes basis functions per image pair jointly with the image transformation. The pipeline requires specifying stage configurations and registration hyperparameters. Similar to section 4.2, we use their default released stage configuration for LuMIR and the stage configuration of their submission to the ReMIND2Reg 2025 challenge (Juvekar et al., 2024).

For ANTs SyN, SynthMorph, and LoCoR, we perform Honkamaa & Marttinen's (2025) hyperparameter optimization protocol based on the HEBO Bayesian optimizer (Cowen-Rivers et al., 2022). Over each validation set, parameters are searched for 50 sampler trials plus 3 repeated runs with different random seeds for the top 3 Gaussian process model candidates.