

Structured SIR: Efficient and Expressive Importance-Weighted Inference for High-Dimensional Image Registration

Ivor J. A. Simpson¹  and Neill D. F. Campbell^{2,3} 

¹ Department of Informatics, University of Sussex, UK
`i.simpson@sussex.ac.uk`

² Department of Computer Science, University College London, UK

³ Department of Computer Science, University of Bath, UK
`neill.campbell@ucl.ac.uk`

Abstract. Image registration is an ill-posed dense vision task, where multiple solutions achieve similar loss values, motivating probabilistic inference. Variational inference has previously been employed to capture these distributions, however restrictive assumptions about the posterior form can lead to poor characterisation, overconfidence and low-quality samples. More flexible posteriors are typically bottlenecked by the complexity of high-dimensional covariance matrices required for dense 3D image registration.

In this work, we present a memory and computationally efficient inference method, Structured SIR, that enables expressive, multi-modal, characterisation of uncertainty with high quality samples. We propose the use of a Sampled Importance Resampling (SIR) algorithm with a novel memory-efficient high-dimensional covariance parameterisation as the sum of a low-rank covariance and a sparse, spatially structured Cholesky precision factor. This structure enables capturing complex spatial correlations while remaining computationally tractable.

We evaluate the efficacy of this approach in 3D dense image registration of brain MRI data, which is a very high-dimensional problem. We demonstrate that our proposed method produces uncertainty estimates that are significantly better calibrated than those produced by variational methods, achieving equivalent or better accuracy. Crucially, we show that the model yields highly structured multi-modal posterior distributions, enable effective and efficient uncertainty quantification.

Keywords: Image Registration · Importance Sampling · Uncertainty Quantification

1 Introduction

Dense image registration, establishing pixel-, or voxel-wise geometric correspondences between images, is a fundamental task in computer vision and medical image analysis. There is intrinsic ambiguity in whether or not images are matched

so this problem is fundamentally ill-posed; there will be multiple plausible registration results. This has led to the use of probabilistic inference approaches both without [16, 24, 28] and with the assistance of neural networks [9, 12, 27]. The use of deep learning has provided some benefits for medical image registration, particularly with respect to computational efficiency via amortised inference [9]. Nonetheless, the expressivity of the posterior distributions provided by existing approaches is limited by the memory scaling of high-dimensional covariance matrices. Moreover, many approaches follow a variational paradigm with the strong assumption that the posterior is an (often factorised) Gaussian, making it impossible to capture multi-modal outputs.

In this work, we propose a framework that relaxes the limitations of variational inference and simplistic posterior distributions. Our contributions include:

1. Introducing an amortised Importance Sampling procedure, Sampled Importance Resampling (SIR), to produce a *proposal distribution*, from which samples can be efficiently drawn and weighted, to establish a more flexible posterior than variational inference that matches the true posterior.
2. Providing an expressive Gaussian formulation for our proposal distribution, with a novel memory-efficient covariance parameterisation, consisting of a sum of a low-rank covariance and a spatially structured locally sparse Cholesky factor of the precision, with cross-directional correlations.
3. Demonstrating the improvements of our inferred importance weighted registration distributions, when compared to variational equivalents, in terms of: (i) segmentation accuracy; (ii) quality and coherence of samples; and, importantly, (iii) probabilistic calibration (i.e. the predicted uncertainty correlates with registration error) on a brain MRI dataset.

2 Background

2.1 Registration Notation

We denote the fixed and moving images as $\mathbf{I}_f$ and $\mathbf{I}_m$, defined over a common discrete spatial domain Ω containing N_v voxels; that is, we are learning the deformation that takes $\mathbf{I}_m$ to $\mathbf{I}_f$. The dense displacement field is stacked in matrix $\mathbf{Z} \in \mathbb{R}^{3 \times N_v}$, where any column $\mathbf{z} \in \mathbb{R}^3$ is a 3D displacement. The warped image is given by $\mathbf{I}_m(\mathbf{x} + \mathbf{z}(\mathbf{x}))$, denoted as $\mathbf{I}_m \circ \mathbf{Z}$ for brevity.

2.2 Probabilistic Registration

While traditional registration provides a single MAP (Maximum-A-Posteriori) estimate, e.g. [1], fully probabilistic frameworks aim to characterise the posterior distribution $p(\mathbf{Z}|\mathbf{I}_f, \mathbf{I}_m)$. This is particularly critical for medical applications, where confidence intervals can guide clinical decision making. However, Uncertainty Quantification (UQ) has been challenging for registration, in general, and for many approaches the characterised uncertainty has been found to correlate poorly with the needs of downstream applications [17]; we seek to rectify this situation with well calibrated uncertainty and coherent posterior samples.

Sampling and Approximate Inference: Previous works attempt to capture the posterior distribution using either: sampling based methods [24, 32], which tend to be slow for dense 3D registration; or variational [16, 28], or Laplace approximation [31], approaches that are both computationally intensive and only provide a unimodal approximation of the posterior.

More recently, Grzech *et al.* [12] bridged the gap between deep learning and classical inference by employing Stochastic Gradient MCMC. While this allows for more robust posterior coverage than standard variational methods, the requirement to draw multiple samples at inference time maintains a significant computational overhead compared to single-pass amortized models.

Amortised Variational Inference: Amortised methods were introduced to resolve the computational expense of estimating an approximate posterior distribution [9]. In this framework, a neural network is trained to predict directly the parameters of an approximate posterior $q_\theta(\mathbf{Z}|\mathbf{I}_f, \mathbf{I}_m)$; typically a multivariate Gaussian. This is achieved by maximizing the Evidence Lower Bound (ELBO)

$$\mathcal{L}_{\text{elbo}} = \mathbb{E}_{q_\theta(\mathbf{Z}|\mathbf{I}_f, \mathbf{I}_m)} [\log p(\mathbf{I}_f|\mathbf{I}_m \circ \mathbf{Z})] + \text{KL}[q_\theta(\mathbf{Z}|\mathbf{I}_f, \mathbf{I}_m)||p(\mathbf{Z})]. \quad (1)$$

Uncertainty Calibration: Although amortised variational methods are efficient, their resulting uncertainty estimates are often poorly calibrated, underestimating the posterior variance [7] and fail to capture multi-modal distributions by construction. To address the multi-scale nature of registration uncertainty, PULPo [27] introduced a hierarchical Laplacian pyramid approach to capture uncertainty across multiple spatial scales. While PULPo improves UQ quality, its iterative/hierarchical nature adds complexity to the sampling procedure.

2.3 Importance Sampling

Amortised importance sampling based inference was first introduced in the Importance Weighted Auto Encoder [4]; a neural network was used to predict a “sample proposal distribution”, rather than an approximate posterior, where samples are weighted at test time based on their likelihood, prior and proposal probabilities.

Importance distributions can be estimated in the same manner as variational methods, where we instead optimise the importance weighting

$$\mathbb{E}_{z^{1:K}} \left[\log \left(\frac{1}{K} \sum_{i=1}^K \frac{p(\mathbf{I}_f, \mathbf{I}_m \circ \mathbf{Z}^i)}{q_\theta(\mathbf{Z}^i|\mathbf{I}_f, \mathbf{I}_m)} \right) \right] = \mathbb{E}_{z^{1:K}} \left[\log \left(\frac{1}{K} \sum_{i=1}^K \alpha^i \right) \right], \quad (2)$$

where

$$\alpha^i = \frac{p(\mathbf{I}_f|\mathbf{I}_m \circ \mathbf{Z}^i)p(\mathbf{Z}^i)}{q_\theta(\mathbf{Z}^i|\mathbf{I}_f, \mathbf{I}_m)} \quad (3)$$

are the unnormalised weights; superscripts, e.g. $\cdot^i$, denote samples.

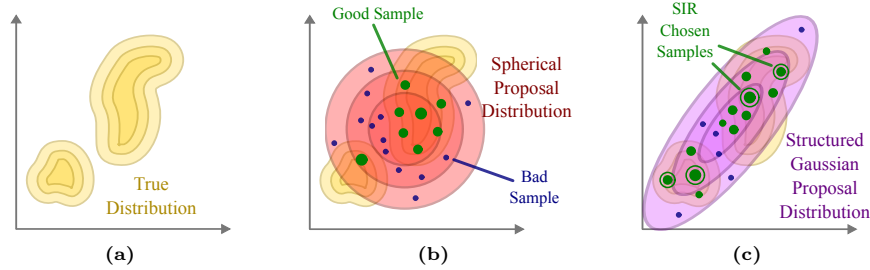


Fig. 1: Illustration of the posterior modelling problem, where the underlying probability distribution is illustrated in (a). (b) shows an attempt to model this variationally using a spherical Gaussian, which assigns probability mass to a lot of area that is unlikely, but simultaneously does not cover the whole distribution; if used as a proposal distribution then many poor samples are drawn and this is exacerbated as dimensionality increases. (c) illustrates how our SIR sampling framework allows selection of high probability samples at both learning/inference time, which avoids penalising a model for being more exploratory. Moreover, using a structured Gaussian allows assigning probability mass to higher density areas improving the efficiency of sampling.

The key distinction from variational inference is that we take the sum of the probabilities, rather than log probabilities, allowing better samples z^i to dominate in the loss (and therefore gradients). The importance log probabilities are thus normalised by the other drawn samples, i.e. assigned importance weights. We note that using a single sample to calculate the expectation is equivalent to variational inference. Previous work demonstrated the multi-modal nature of the inferred distribution [8] but, to the best of our knowledge, this has not been directly applied to dense imaging problems.

2.4 Flexible Covariances

Structured Uncertainty Prediction Networks [10], demonstrated an efficient amortised inference approximation to a full-rank covariance via a sparse parameterisation of the *precision* (the inverse of the covariance) equivalent to a Gaussian Markov Random Field [26]. Low-rank covariance approximations have also been shown to be effectively learned using amortised inference [20]. In this work, we integrate both these formulations together into a more general and flexible parameterisation.

3 Methods

3.1 Sampled Importance Resampling (SIR)

To approximate the complex posterior $p(\mathbf{Z}|\mathbf{I}_f, \mathbf{I}_m)$, we employ the SIR framework [25, 29], a form of importance weighted inference that approximates the objective in Eq. (2). While traditional approaches use all the samples from a

proposal distribution to compute a weighted average, we draw inspiration from the resampling-based objectives in [18], which focuses the gradient computation on a subset of high-quality samples.

Specifically, we use SIR to identify high-probability registration hypotheses from a large set of candidates drawn from the initial proposal distribution; these selected samples are then used to provide an informative supervisory signal for optimising the parameters of our proposal distribution network. This reduces the overhead of backpropagating through low-weight candidates, and permits a more expansive search of the proposal space. We draw N_s samples from the learned proposal distribution $q(\mathbf{Z}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ using the reparameterisation trick

$$\mathbf{Z}^i = \boldsymbol{\mu} + \mathbf{R}\boldsymbol{\epsilon}_R^i + \mathbf{L}^{-\top}\boldsymbol{\epsilon}_L^i, \quad (4)$$

where $\boldsymbol{\epsilon}_L$ and $\boldsymbol{\epsilon}_R$ are samples from a standard Normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{R}$ and $\mathbf{L}$ are defined in Sec. 3.2 as part of our covariance parameterisation.

As we are modelling a high-dimensional correlated distribution $\mathbf{Z} \in \mathbb{R}^{3 \times N_v}$, we might need to draw many samples to find effective ones, which would be computationally expensive and prohibitive in terms of memory to backpropagate through lots of low weighted samples. As such, we draw these samples without gradients but store the noise components for use in a subsequent backwards pass through selected samples.

To identify the best samples to focus on, we compute our unnormalized importance weights without gradients, α^i , based on the ratio between the target posterior and the proposal density

$$\alpha^i = \frac{p(\mathbf{I}_f | \mathbf{I}_m \circ \mathbf{Z}^i) p(\mathbf{Z}^i)}{q(\mathbf{Z}^i)}. \quad (5)$$

The normalized weights are then defined as $w^i = \alpha^i / \sum_{j=1}^{N_s} \alpha^j$. These weights represent the relative “quality” of each registration hypothesis under the image evidence.

To compute gradients for the network parameters, we form a weighted estimate of the posterior. Rather than backpropagating through all N_s samples, many of which may have negligible weights, we perform multinomial resampling of these samples based on the normalised weight vector $\mathbf{w}$. We draw N_k indices, and then use these selected sample random noise vectors ($\boldsymbol{\epsilon}_R^k$ and $\boldsymbol{\epsilon}_L^k$) to re-draw the samples and compute α^k , but this time enabling gradients to flow back through the network to update the model weights. As we have now selected samples based on importance weighting, our final loss is simply the expected negative log-likelihood over these samples, that is

$$\mathcal{L}_{\text{SIR}} = -\frac{1}{N_k} \sum_{k=1}^{N_k} (\log p(\mathbf{I}_f | \mathbf{I}_m \circ \mathbf{Z}^k) + \log p(\mathbf{Z}^k)). \quad (6)$$

Likelihood Function Similarly to [2] we use normalised cross-correlation as our likelihood model, which provides robustness against global intensity fluctuations

common in MRI acquisition. We treat it as the energy function of a Boltzmann distribution:

$$p(\mathbf{I}_f | \mathbf{I}_m, \mathbf{Z}) \propto \exp \left(\frac{-\text{NCC}(\mathbf{I}_f, \mathbf{I}_m \circ \mathbf{Z})}{\sigma^2} \right), \quad (7)$$

where σ^2 is a likelihood scaling function.

Prior Function For simplicity we use a diffusion based prior, where we penalise the squared spatial gradients of the deformation field:

$$p(\mathbf{Z}) \propto \exp \left(-\frac{\lambda}{2} \sum_{j=1}^{N_v} \|\nabla \mathbf{z}_j\|^2 \right) + \left(-\frac{\lambda_{\boldsymbol{\mu}}}{2} \sum_{j=1}^{N_v} \|\nabla \boldsymbol{\mu}_j\|^2 \right). \quad (8)$$

We employ a stronger regularisation weight $\lambda_{\boldsymbol{\mu}}$ on $\boldsymbol{\mu}$ than on the sampled field $\mathbf{Z}$ to allow the model to explore a wider range of the displacement manifold. This prevents the distribution becoming overly stiff, and enables the resulting samples to capture the multi-modal nature of the posterior. For fairness, we use the same prior in our variational experiments as well.

3.2 Covariance Parameterisation

We define the covariance matrix for our proposal, or variational, distribution as a sum of a low-rank and a sparse structured Cholesky factored matrix: $\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}^\top + (\mathbf{L}\mathbf{L}^\top)^{-1}$ where $\mathbf{R} \in \mathbb{R}^{N_v \times r}$ and $\mathbf{L} \in \mathbb{R}^{N_v \times N_v}$ is a sparse lower triangular matrix with positive diagonal entries. In contrast to the 2D SUPN structure of [10], our sparsity structure for $\mathbf{L}$ runs over the 3D spatial structure of the voxel grid and over the three channels of the output displacement vector.

Drawing samples from this is efficient, as we can draw from each covariance component independently, as noted in Eq. (4). We use the Cholespy solver [23] for efficient sampling using $\mathbf{L}^{-\top}$.

To enable importance weighting, we need to evaluate the log likelihood under the proposal distribution of any sample $\mathbf{Z}$, which is expressed as

$$\log p(\mathbf{Z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \left(\underbrace{(\text{vec}(\mathbf{Z}) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\text{vec}(\mathbf{Z}) - \boldsymbol{\mu})}_{\boldsymbol{\Sigma}^{-1} \text{ quadratic prod}} + \underbrace{\log \det(2\pi \boldsymbol{\Sigma})}_{\log \det} \right), \quad (9)$$

where the displacement field is vectorised, $\text{vec}(\mathbf{Z}) \in \mathbb{R}^{3N_v}$. The quadratic inverse covariance product can be evaluated, by applying the Woodbury Matrix Inversion Lemma, as

$$(\text{vec}(\mathbf{Z}) - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\text{vec}(\mathbf{Z}) - \boldsymbol{\mu}) = \mathbf{k}^\top \mathbf{k} - (\mathbf{S}^\top \mathbf{k})^\top \mathbf{M}^{-1} (\mathbf{S}^\top \mathbf{k}), \quad (10)$$

where $\mathbf{k} = \mathbf{L}^\top (\mathbf{Z} - \boldsymbol{\mu})$, $\mathbf{S} = \mathbf{L}^\top \mathbf{R}$ and $\mathbf{M} = \mathbf{I}^{-1} + \mathbf{S}^\top \mathbf{S}$, and $\mathbf{M} \in \mathbb{R}^{r \times r}$, which we invert using an eigenvalue decomposition (to avoid numerical instabilities). The log determinant term can be evaluated as

$$\log \det \boldsymbol{\Sigma} = \log \det((\mathbf{L}\mathbf{L}^\top)^{-1} + \mathbf{R}\mathbf{R}^\top) = -\log \det(\mathbf{L}\mathbf{L}^\top) + \log \det(\mathbf{M}). \quad (11)$$

A full derivation is given in Appendix A.

3.3 Network Architecture

We use a fully convolutional U-net style architecture, without bias terms. The model has 4 resolution levels, with 16 channels for the first resolution level of the encoder, and 24 channels thereafter. We have separate convolutional prediction heads for the mean, $\boldsymbol{\mu}$, the sparse Cholesky factor $\mathbf{L}$, and the low-rank matrices $\mathbf{R}$. The first two are calculated at half the voxel resolution (to improve efficiency) and upsampled using a cubic spline, for C^2 smoothness; the low-rank matrices are produced at half that resolution level, to reduce memory demands, and upsampled to match the Cholesky factor using bilinear interpolation. No activation functions are applied to $\boldsymbol{\mu}$ or $\mathbf{R}$; for $\mathbf{L}$ we use a soft-plus for the Cholesky diagonals, to ensure they are positive, and no activation for the off-diagonal terms.

4 Experiments

4.1 Implementation Details

All models were trained for 50,000 steps with a batch size of 1 on a single NVIDIA RTX A6000 (48 GB VRAM), with a peak memory requirement of 32 GB. We utilized the SOAP optimizer [30] via the Heavyball library [22], which facilitates stable training at high learning rates; we set the learning rate to 2×10^{-3} with $\beta_1 = 0.9$ and $\beta_2 = 0.95$. A cosine learning rate scheduler was employed, decaying to 2×10^{-4} after 30,000 iterations. To improve model robustness, we applied data augmentation consisting of coherent voxel shifts (up to 8 voxels) and independent (to each image in the pair) voxel shifts (up to 2 voxels).

4.2 Hyperparameters

We fix the number of low-rank components to 25, unless noted otherwise and provide additional results in Appendix F.1. The Cholesky connectivity pattern $\mathbf{L}$ accounts for neighbors within a $3 \times 3 \times 3$ kernel with cross-channel correlations between deformation directions, larger neighbours would require too much VRAM. We apply L_2 regularization with a coefficient of 0.05 to the off-diagonal elements of $\mathbf{L}$ to prevent overfitting and ensure numerical stability during inference.

We use a ratio of $\lambda_{\boldsymbol{\mu}} = 2.5$, $\lambda = 1.0$ and tuned the likelihood scale parameter to $\sigma = 0.5$ on the variational baseline for segmentation accuracy while maintaining less than 0.05% folding of the deformation field.

For the SIR-based variants, we introduce a temperature scaling hyperparameter to prevent the importance weight distribution becoming too spiky (dominated by individual samples), or flat (uniform sampling). We adopt a dynamic temperature scaling approach [21], where the unnormalised weights $\boldsymbol{\alpha}$ are scaled by a factor of $\frac{T}{\sigma_{\boldsymbol{\alpha}}}$. Here $\sigma_{\boldsymbol{\alpha}}$ is an exponential moving average of standard deviation of the weights, i.e. $\sigma_{\boldsymbol{\alpha}}^t = \gamma \sigma_{\boldsymbol{\alpha}}^{t-1} + (1 - \gamma) \text{std}(\boldsymbol{\alpha})$, where t indicates the optimisation step and $\gamma = 0.9$. We select $T = 3.0$, which empirically provides

a trade-off between uniform, and highly selective weights. Further temperature experiments are in Appendix F.2.

4.3 Inference and Posterior Characterisation

At inference time, we draw $N_l = 1,200$ samples (in batches), and perform importance weighting on these. Through multinomial sampling, we then draw a total of $N_k = 80$ samples to characterise the posterior distribution. While we use the final EMA value of σ_α obtained from training for test-time scaling, we note that this could be optimised post-training to improve performance.

4.4 Model Variants

We evaluate the performance of our model in terms of propagated segmentation accuracy after registering pairs of images. We denote the baseline variational variants of our model with a leading **V**, regular importance sampling has a leading **I**, and SIR with **S**. Symbols after the $+$ represent the covariance parameterisation: **D** is a diagonal Cholesky, **C** is a sparsely structured Cholesky and **L** indicates a low-rank component. We also compare against a resolution matched Voxelmorph model [9], with a comparable architecture (see Appendix G for details) that we ensembled using 5 models (Vxm5).

4.5 Data

We use a subset of the OASIS [19] as pre-processed and described in [2]. Each image has 35 automatically segmented labels using [11], 17 of which are bilateral, and is resampled to 1 mm^3 isotropic voxels with an image size of $160 \times 192 \times 224$. The resolution of the predicted grid, following cropping some of the background, is $96 \times 72 \times 80$, leading to $N_v = 552,960$ and considering a 3D displacements field leads to us estimating a multivariate distribution over 1.6 million dimensions.

The data contains substantial anatomical variations and differences in brain shape. It also has dense high-resolution anatomical labels of varying scales and morphological complexity, enabling us to validate the uncertainty estimates that arise. During both training and evaluation, we perform subject-to-subject registration, where each image is registered to others within its respective set. For the test set, this results in $50 \times 49 = 2,450$ unique registration pairs to ensure a rigorous assessment of model generalisation and uncertainty calibration.

5 Results

5.1 Accuracy

Fig. 2 illustrates the improvements in the Dice Similarity Coefficient (DSC) for specific structures when using S+LC over V+LC. Most strikingly, we observe a substantial selection gap, which we define as the improvement in the Oracle

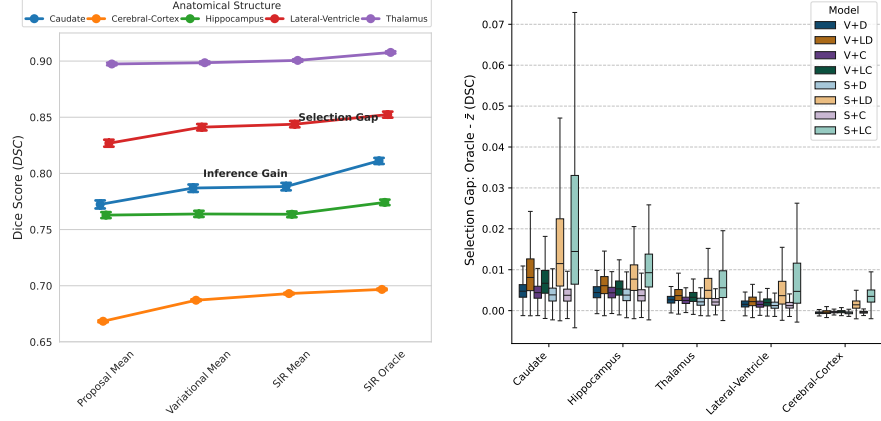


Fig. 2: Left, an illustration of the proposal mean (μ in SIR), variational μ , SIR sample mean and Oracle (best-sample), and standard deviation of the segmentation accuracy for different model variants. Right, a boxplot exemplifying the quality of the Oracle sample drawn from the approximate posterior distribution. Larger Dice Similarity Coefficient (DSC) values indicated greater segmentation accuracy.

Table 1: Registration performance and folding rates across OASIS test pairs. We report the Dice Similarity Coefficient (DSC) for the mean prediction ($\mu_{\mathbf{Z}}$), the mean of resampled fields ($\bar{\mathbf{Z}}$), and the Oracle (the single sample from the posterior with the highest DSC). Fold % represents the percentage of voxels with a non-positive Jacobian determinant. Standard deviations are shown in parentheses. High DSC values and lower fold % are preferable.

Model	μ (DSC)	$\bar{\mathbf{Z}}$ (DSC)	Oracle (DSC)	μ fold %	$\bar{\mathbf{Z}}$ fold %	Oracle fold %
V+D	0.800 (0.04)	0.800 (0.04)	0.805 (0.04)	0.046 (0.04)	0.059 (0.05)	0.059 (0.05)
V+LD	0.799 (0.04)	0.798 (0.04)	0.805 (0.04)	0.046 (0.04)	0.059 (0.05)	0.059 (0.05)
V+C	0.802 (0.04)	0.802 (0.04)	0.807 (0.04)	0.045 (0.05)	0.057 (0.06)	0.057 (0.06)
V+LC	0.800 (0.04)	0.800 (0.04)	0.806 (0.04)	0.044 (0.03)	0.056 (0.04)	0.056 (0.04)
I+LC	0.800 (0.04)	0.800 (0.04)	0.805 (0.04)	0.050 (0.05)	0.063 (0.06)	0.063 (0.06)
S+D	0.801 (0.04)	0.801 (0.04)	0.805 (0.04)	0.048 (0.05)	0.062 (0.06)	0.062 (0.06)
S+LD	0.798 (0.04)	0.800 (0.04)	0.809 (0.04)	0.032 (0.03)	0.085 (0.08)	0.096 (0.10)
S+C	0.801 (0.04)	0.801 (0.04)	0.805 (0.04)	0.051 (0.03)	0.065 (0.04)	0.065 (0.04)
S+LC	0.794 (0.04)	0.801 (0.04)	0.812 (0.03)	0.011 (0.01)	0.115 (0.07)	0.163 (0.13)
S+LC 10L	0.788 (0.04)	0.802 (0.04)	0.814 (0.03)	0.004 (0.01)	0.240 (0.14)	0.292 (0.25)
Vxm5	0.798 (0.03)	0.798 (0.03)	0.804 (0.03)	0.000 (0.00)	0.000 (0.00)	0.000 (0.00)

(most accurate) segmentation performance, compared to the SIR or Variational mean. This is further examined in Tab. 1, where we summarise the average segmentation propagation accuracies and observe all approaches have similar average performance but S+LC and S+LD show the biggest gains. A complete set of boxplots for different structures is provided in Appendix B. The more complex posterior forms offer limited benefits in variational inference but confer a pronounced effect in SIR models. In our sensitivity analysis (Appendix F.1), we found that 10 low-rank components outperforms in 25 in terms of accuracy, and calibration, while also being more efficient to sample. We also observe a slight improvement in accuracy over the 5 model Voxelmorph ensemble, at the expense of folding. We note that the level of voxel folding increases with the more exploratory, high-variance deformations found by SIR, which are particularly noticeable in the Oracle samples. Further evaluation of the oracle samples is given in Appendix C.

5.2 Calibration of the Posterior

Table 2: Uncertainty Calibration Metrics: Area Under Sparse Error (AUSE) based on label entropy and Expected Calibration Error (ECE) calculated using a dilated (by 3 voxels) segmentation mask. The last column is the Spearman correlation of label entropy and Dice. All fields show the average and std-dev over structures.

Model	AUSE _{label} ↓	ECE ↓	Label Entropy	Mean Spearman r: Entropy & DSC
V+D	0.0850 (0.021)	0.1285 (0.027)	0.0628 (0.002)	0.064 (0.169)
V+LD	0.0818 (0.022)	0.1257 (0.027)	0.0705 (0.002)	0.085 (0.165)
V+C	0.0854 (0.021)	0.1279 (0.026)	0.0595 (0.002)	0.046 (0.164)
V+LC	0.0844 (0.021)	0.1274 (0.027)	0.0641 (0.002)	0.033 (0.167)
I+LC	0.0861 (0.021)	0.1287 (0.027)	0.0621 (0.002)	0.011 (0.172)
S+D	0.0897 (0.021)	0.1312 (0.026)	0.0549 (0.007)	0.051 (0.123)
S+LD	0.0736 (0.020)	0.1159 (0.026)	0.0901 (0.004)	-0.137 (0.177)
S+C	0.0907 (0.021)	0.1324 (0.027)	0.0531 (0.006)	0.007 (0.121)
S+LC	0.0659 (0.019)	0.1064 (0.025)	0.1079 (0.006)	-0.325 (0.218)
S+LC 10L	0.0636 (0.018)	0.1028 (0.023)	0.1156 (0.010)	-0.419 (0.217)
Vxm5	0.1008 (0.015)	0.1224 (0.020)	0.0703 (0.007)	-0.410 (0.209)

We evaluate calibration using both the Expected Calibration Error (ECE) and Area Under the Sparsification Error (AUSE) for each segmentation region. To focus our metric on relevant regions, we compute ECE and AUSE within a 3-voxel dilated boundary of the anatomical labels, as registration ambiguity and label uncertainty are most pronounced at tissue interfaces, specific details given in Appendix H. ECE quantifies the average discrepancy between a model’s

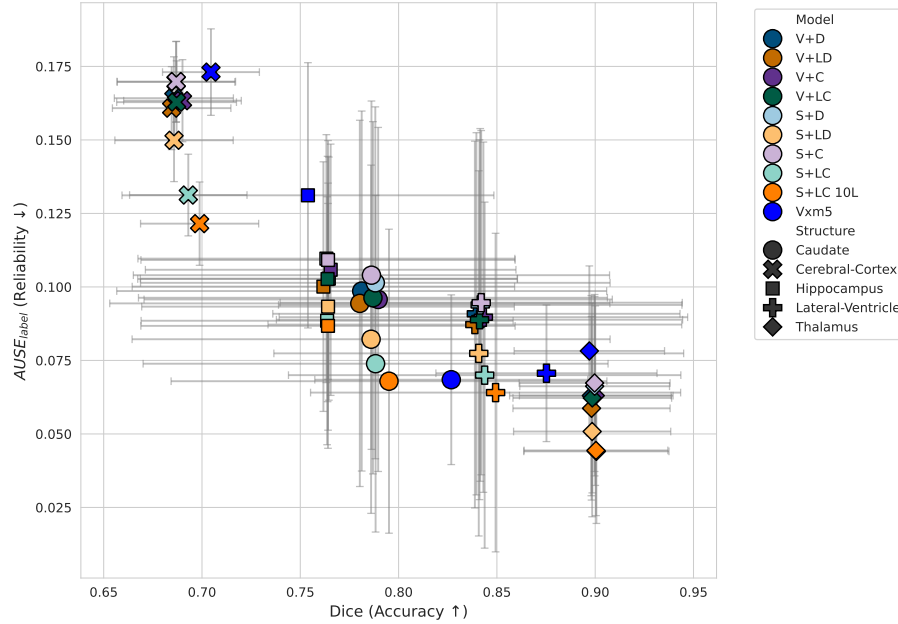


Fig. 3: Accuracy vs calibration for the different model variants for a set of structures, standard deviations plotted in gray. The best performance is in the bottom-right of the graph with a low AUSE (better calibration) and a high DSC (better accuracy). The full version of the model (S+LC) with 10 low-rank components offers improved calibration with equal or improved accuracy consistently over all structures. The 5 model Voxelmorph ensemble has improved average accuracy for some structures, but generally much worse calibration.

predicted probability of a label and the actual observed frequency of that label; in the context of registration, a lower ECE indicates that the model’s confidence scores are more reliable indicators of anatomical alignment. Whereas, AUSE evaluates uncertainty by measuring how registration error decreases as we progressively remove voxels with the highest uncertainty; a lower AUSE signifies that the uncertainty map is an effective proxy for localized registration errors

We evaluate the calibration of our models in Tab. 2. We find that our SIR models with low-rank covariance parameterisations, particularly S+LC with 25 or 10 low-rank components, are much better calibrated than the variational models, the naive importance sampling approach (which performs similarly to variational inference), and the Voxelmorph deep ensemble in terms of both ECE and AUSE. We provide a structure-wise evaluation in Appendix D. Furthermore, we examined the correlation between label entropy and accuracy (in terms of DSC) for each structure, and averaged these. We find that the poorly calibrated models exhibit small **positive** correlations, incorrectly indicating more uncertainty when they are most accurate. Whereas, our better calibrated models, such

as S+LC and S+LC 10L, demonstrate strong negative correlations indicating the uncertainty is a much more reliable proxy for registration accuracy.

We highlight the relationship between accuracy and calibration for the different model variants in Fig 3, which clearly illustrates that S+LC leads to improved calibration without any implications on accuracy.

5.3 Multi-Modality

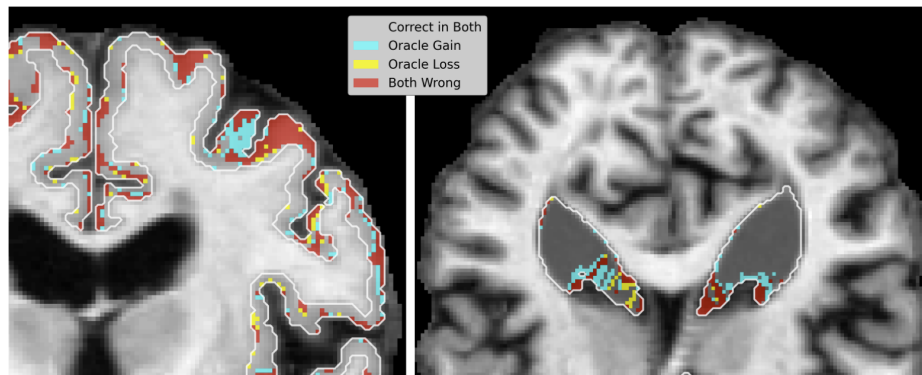


Fig. 4: Illustration of the structured nature of the samples, where we find the performance of the Oracle sample can lead to substantial structured gains over the average SIR deformation. On the left, we show an example where a lobe of the cerebral cortex is correctly aligned by the oracle sample, with a DSC improvement of 0.04. On the right, we observe substantial structured changes in the lateral ventricle, with a DSC improvements of 0.03.

The spatial coherence, and multi-modality, of the learned distribution is exemplified by examining the improvement in segmentation performance by the Oracle sample. As shown in Fig. 4, we can see segmentation corrections in smooth groups of pixels, which is enabled through our learned spatial distribution.

To inspect individual cases, we performed PCA on the SIR sample deformations, run clustering to identify groups (in our case the top two) and then picked examples nearest to those cluster means. Two examples of bimodality with equivalent DSC accuracy are given in Fig. 5.

6 Discussion

Our results demonstrate that our proposed SIR approach, when coupled with structured covariance parameterisations, significantly improves the calibration of registration uncertainty without compromising accuracy. By moving beyond the restrictive assumptions of amortised variational inference, we demonstrate

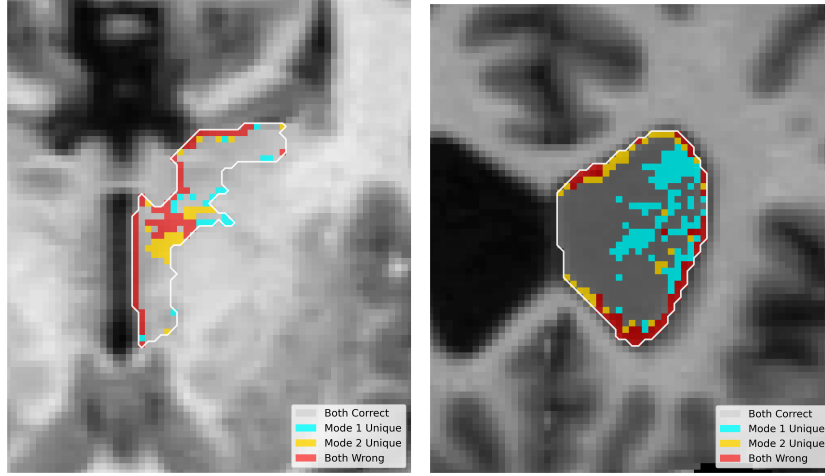


Fig. 5: Illustration of the multi-modality of the posterior distribution, where disagreements between modes are marked in blue or orange. On the left, we find structured changes in the Thalamus where mode 1 has a DSC of 0.834 and mode 2 has a Dice of 0.824. On the right, we observe substantial differences in the left lateral ventricle, either on the interior or exterior border, with a DSC of 0.864 for mode 1 and 0.862 for mode 2.

that registration uncertainty can be characterised in a manner that serves as a reliable proxy for anatomical alignment.

The SIR framework acts as a consistent surrogate to the importance weighted objective in Eq. (2), analogous to the Reweighted Wake-Sleep ‘wake’ update [3], stabilizing training by decoupling sample weighting from gradient propagation. While not a theoretically guaranteed optimizer of Eq. (2), our empirical results and their stability over training Appendix E, demonstrate that this strategy effectively guides the proposal distribution toward high-probability regions while avoiding the entropy collapse seen in direct importance weighted optimisation.

6.1 Calibration and Meaningful Uncertainty

A central finding in this work relates to the relationship between model uncertainty and segmentation error. Our proposed S+LC model significantly outperforms variational baselines and Voxelmorph ensembles in terms of calibration metrics, but also exhibits a strong negative correlation between label entropy and accuracy. The expanded entropy range of the proposed model, coupled with its superior registration accuracy, suggests that the captured uncertainty is highly structured within the high-dimensional deformation space, representing a coherent distribution of anatomically plausible hypotheses rather than stochastic noise.

6.2 The Accuracy of Oracle Samples

The selection gap between the SIR mean, and the Oracle provides an intuition into the complexity of the inferred posterior. It indicates that very good samples exist with high-probability in the proposal distribution, but perhaps due to multi-modality do not coincide with the mean. This finding supports the use of SIR for fitting complex posterior distributions.

6.3 Computational Efficiency

Despite the high dimension of the displacement field, our approach remains computationally efficient, requiring a single forward pass of the amortised model to produce an effective proposal distribution from which samples are drawn. Samples require a sparse linear solve for $\mathbf{L}$, which is $O(\text{nnz}(\mathbf{L}))$, and a matrix multiplication for $\mathbf{R}$. The entire process of proposal inference, sample drawing and weighting of 1,200 samples is completed in under 5 seconds on a single GPU.

This represents a substantial computational benefit over hierarchical sampling approaches such as PULPo [27], which require sequential forward passes through the model for each sample, coupling the sampling complexity with the depth of the neural network, or MCMC-based methods [12], which require multiple iterative passes.

6.4 Limitations and Future Directions

While our approach improves calibration, with good accuracy, there is a trade-off in terms of topological regularity and folding of voxels. Future work could investigate using stationary velocity fields, as in [9], to encourage diffeomorphic transformations. Furthermore, more complex regularisers [5] may offer benefits in certain applications.

Our architecture was chosen based on our existing code library, however the underlying SIR and covariance parameterisation are model-agnostic. Future directions could replace the convolutional backbone with a vision transformer architecture [6].

In terms of improvement to segmentation calibration, we could investigate post-training optimisation of the temperature parameter, and more rigorous hyper-parameter selection. We could also include segmentation losses within the likelihood function; although these are inaccessible at test time, but they could contribute to learning the shape of the importance sampling distribution.

Beyond medical image registration, our proposed integration of SIR with our structured sparsity and low-rank covariance model offers a versatile framework for high-dimensional spatial inference tasks in inverse problems that suffer from ambiguity, such as optical flow [15], sparse-view CT or 3D image reconstruction and image super-resolution.

7 Conclusion

We have introduced Structured SIR, a framework for efficient and expressive approximate inference of posterior distributions in high-dimensional imaging tasks. We have demonstrated the application of this to probabilistic image registration, where we leverage SIR with a novel structured covariance parameterisation to characterise a 1.6 million dimensional posterior distribution of a 3D deformation field.

Our proposed SIR algorithm makes use of a dual sampling pass: first evaluating a large set of candidates to weight based on their respective probabilities, then performing resampling of these to characterise the posterior. To the best of our knowledge, this is the first application of SIR to a dense vision task of this scale.

Our results demonstrate that SIR with sufficiently complex covariance parameterisations, exhibits good calibration performance, substantially outperforming variational inference while maintaining accuracy. Moreover, the substantial gap in the accuracy of the Oracle sample, confirms that our expressive structured proposal distribution supports a diverse range of plausible deformations.

This framework offers a scalable blueprint for uncertainty quantification in high-dimensional spatial tasks. While demonstrated here for registration, it could be extensible to other inverse problems where test-time evidence can guide the selection of solutions.

Acknowledgements

NC acknowledges support from the EPSRC CAMERA Research Centre (EP/T022523/1), the UKRI Strength in Places Fund My-World Project (SIPF00006/1) and the Royal Society. We thank Teo Deveney and Paula Seidler for their help refactoring the supporting codebase for Cholesky calculation and sampling.

References

1. Ashburner, J.: A fast diffeomorphic image registration algorithm. *Neuroimage* **38**(1), 95–113 (2007)
2. Balakrishnan, G., Zhao, A., Sabuncu, M.R., Guttag, J., Dalca, A.V.: Voxelmorph: a learning framework for deformable medical image registration. *IEEE transactions on medical imaging* **38**(8), 1788–1800 (2019)
3. Bornschein, J., Bengio, Y.: Reweighted wake-sleep. *arXiv preprint arXiv:1406.2751* (2014)
4. Burda, Y., Grosse, R., Salakhutdinov, R.: Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519* (2015)
5. Burger, M., Modersitzki, J., Ruthotto, L.: A hyperelastic regularization energy for image registration. *SIAM Journal on Scientific Computing* **35**(1), B132–B148 (2013)

6. Chen, J., Frey, E.C., He, Y., Segars, W.P., Li, Y., Du, Y.: Transmorph: Transformer for unsupervised medical image registration. *Medical image analysis* **82**, 102615 (2022)
7. Cremer, C., Li, X., Duvenaud, D.: Inference suboptimality in variational autoencoders. In: *International conference on machine learning*. pp. 1078–1086. PMLR (2018)
8. Cremer, C., Morris, Q., Duvenaud, D.: Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916* (2017)
9. Dalca, A.V., Balakrishnan, G., Guttag, J., Sabuncu, M.R.: Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical image analysis* **57**, 226–236 (2019)
10. Dorta, G., Vicente, S., Agapito, L., Campbell, N.D., Simpson, I.: Structured uncertainty prediction networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5477–5485 (2018)
11. Fischl, B.: Freesurfer. *Neuroimage* **62**(2), 774–781 (2012)
12. Grzech, D., Azampour, M.F., Qiu, H., Glocker, B., Kainz, B., Folgoc, L.L.: Uncertainty quantification in non-rigid image registration via stochastic gradient markov chain monte carlo. *arXiv preprint arXiv:2110.13289* (2021)
13. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical image analysis* **16**(7), 1423–1435 (2012)
14. Heinrich, M.P., Papież, B.W., Schnabel, J.A., Handels, H.: Non-parametric discrete registration with convex optimisation. In: *Biomedical Image Registration: 6th International Workshop, WBIR 2014, London, UK, July 7-8, 2014. Proceedings 6*. pp. 51–61. Springer (2014)
15. Ilg, E., Cicek, O., Galesso, S., Klein, A., Makansi, O., Hutter, F., Brox, T.: Uncertainty estimates and multi-hypotheses networks for optical flow. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 652–667 (2018)
16. Le Folgoc, L., Delingette, H., Criminisi, A., Ayache, N.: Quantifying registration uncertainty with sparse bayesian modelling. *IEEE transactions on medical imaging* **36**(2), 607–617 (2016)
17. Luo, J., Sedghi, A., Popuri, K., Cobzas, D., Zhang, M., Preiswerk, F., Toews, M., Golby, A., Sugiyama, M., Wells III, W.M., et al.: On the applicability of registration uncertainty. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 410–419. Springer (2019)
18. Maddison, C.J., Lawson, J., Tucker, G., Heess, N., Norouzi, M., Mnih, A., Doucet, A., Teh, Y.: Filtering variational objectives. *Advances in neural information processing systems* **30** (2017)
19. Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L.: 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)
20. Monteiro, M., Le Folgoc, L., Coelho de Castro, D., Pawlowski, N., Marques, B., Kamnitsas, K., Van der Wilk, M., Glocker, B.: Stochastic segmentation networks: Modelling spatially correlated aleatoric uncertainty. *Advances in neural information processing systems* **33**, 12756–12767 (2020)
21. Neal, R.M.: Annealed importance sampling. *Statistics and computing* **11**(2), 125–139 (2001)
22. Nestler, L., HomebrewML contributors: HeavyBall (Apr 2026), <https://github.com/HomebrewML/HeavyBall>

23. Nicolet, B., Jacobson, A., Jakob, W.: Large steps in inverse rendering of geometry. *ACM Transactions on Graphics (TOG)* **40**(6), 1–13 (2021)
24. Risholm, P., Golby, A.J., Wells, W.: Multimodal image registration for preoperative planning and image-guided neurosurgical procedures. *Neurosurgery Clinics* **22**(2), 197–206 (2011)
25. Rubin, D.B.: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association* **82**(398), 542–543 (1987)
26. Rue, H., Held, L.: *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC (2005)
27. Siegert, L., Fischer, P., Heinrich, M.P., Baumgartner, C.F.: Pulpo: probabilistic unsupervised laplacian pyramid registration. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 717–727. Springer (2024)
28. Simpson, I.J., Cardoso, M.J., Modat, M., Cash, D.M., Woolrich, M.W., Andersson, J.L., Schnabel, J.A., Ourselin, S., Initiative, A.D.N., et al.: Probabilistic non-linear registration with spatially adaptive regularisation. *Medical image analysis* **26**(1), 203–216 (2015)
29. Smith, A.F., Gelfand, A.E.: Bayesian statistics without tears: a sampling–resampling perspective. *The American Statistician* **46**(2), 84–88 (1992)
30. Vyas, N., Morwani, D., Zhao, R., Kwun, M., Shapira, I., Brandfonbrener, D., Janson, L., Kakade, S.: Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321* (2024)
31. Wang, J., Wells III, W.M., Golland, P., Zhang, M.: Efficient laplace approximation for bayesian registration uncertainty quantification. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 880–888. Springer (2018)
32. Zhang, M., Singh, N., Fletcher, P.T.: Bayesian estimation of regularization and atlas building in diffeomorphic image registration. In: *International conference on information processing in medical imaging*. pp. 37–48. Springer (2013)

A Covariance Formulation

We define our covariance matrix as

$$\boldsymbol{\Sigma} = \mathbf{R}\mathbf{R}^\top + (\mathbf{L}\mathbf{L}^\top)^{-1}, \quad (12)$$

where $\mathbf{R} \in \mathbb{R}^{n \times r}$ and $\mathbf{L} \in \mathbb{R}^{n \times n}$ (but sparse). Samples are easy to obtain, as per Eq. (4), as a draw from each covariance component.

We need to consider the log probability under this density

$$\log p(\text{vec}(\mathbf{Z}); \boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2} \left(\underbrace{\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x}}_{\boldsymbol{\Sigma}^{-1} \text{prod}} + \underbrace{\log \det(2\pi \boldsymbol{\Sigma})}_{\log \det} \right), \quad (13)$$

where we have $\mathbf{x} := \text{vec}(\mathbf{Z}) - \boldsymbol{\mu}$ for generalisation purposes, and to aid readability. We derive a computational tractable form for each term individually in the following sections.

A.1 Inverse Covariance Product

Considering

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} = \mathbf{x}^\top (\mathbf{R}\mathbf{R}^\top + (\mathbf{L}\mathbf{L}^\top)^{-1})^{-1} \mathbf{x}, \quad (14)$$

we can apply the Woodbury Matrix identity

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}, \quad (15)$$

where $\mathbf{A} = (\mathbf{L}\mathbf{L}^\top)^{-1}$, so $\mathbf{A}^{-1} = \mathbf{L}\mathbf{L}^\top$, $\mathbf{V} = \mathbf{U}^\top = \mathbf{R}^\top$ and $\mathbf{C} = \mathbf{I}$. Therefore, we seek to calculate

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} = \mathbf{x}^\top (\mathbf{A} + \mathbf{UCV})^{-1} \mathbf{x} \quad (16)$$

$$= \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} - \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1} \mathbf{VA}^{-1} \mathbf{x}, \quad (17)$$

where $\mathbf{k} := \mathbf{L}^\top \mathbf{x}$, which we can efficiently calculate via convolution operations [10], reducing the first term of the expression to a simple quadratic product $\mathbf{k}^\top \mathbf{k}$.

Turning our attention to the second term, if we denote

$$\mathbf{M} = \mathbf{C}^{-1} + \mathbf{R}^\top \mathbf{A}^{-1} \mathbf{R}, \quad (18)$$

where $\mathbf{M} \in \mathbb{R}^{r \times r}$ and note that

$$\mathbf{VA}^{-1} \mathbf{x} = \mathbf{V}\mathbf{L}\mathbf{L}^\top \mathbf{x} = \mathbf{V}\mathbf{L}\mathbf{k} = \mathbf{S}^\top \mathbf{k}, \quad (19)$$

where $\mathbf{S}^\top := \mathbf{V}\mathbf{L}$. Therefore $\mathbf{S} = \mathbf{L}^\top \mathbf{U}$ that we can again calculate efficiently through convolving $\mathbf{L}^\top$ with the low-rank images in $\mathbf{R}$. We can then write $\mathbf{M} = \mathbf{C}^{-1} + \mathbf{S}^\top \mathbf{S}$.

Putting it all together we have

$$\mathbf{x}^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} = \mathbf{x}^\top (\mathbf{A} + \mathbf{UCV})^{-1} \mathbf{x} \quad (20)$$

$$= \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{x} - \mathbf{x}^\top \mathbf{A}^{-1} \mathbf{U} (\mathbf{C}^{-1} + \mathbf{VA}^{-1} \mathbf{U})^{-1} \mathbf{VA}^{-1} \mathbf{x} \quad (21)$$

$$= \mathbf{k}^\top \mathbf{k} - \mathbf{x}^\top \mathbf{LL}^\top \mathbf{UM}^{-1} \mathbf{VLL}^\top \mathbf{x} \quad (22)$$

$$= \mathbf{k}^\top \mathbf{k} - (\mathbf{k}^\top \mathbf{S})^\top \mathbf{M}^{-1} (\mathbf{S}^\top \mathbf{k}) \quad (23)$$

$$= \mathbf{k}^\top \mathbf{k} - (\mathbf{S}^\top \mathbf{k})^\top \mathbf{M}^{-1} (\mathbf{S}^\top \mathbf{k}). \quad (24)$$

As $\mathbf{M}$ is only of size $r \times r$, where $r \ll n$ (in this paper $r = 25$), calculation of a Cholesky factorisation, or Eigen-decomposition, is computationally tractable. In practice, we use an Eigen-decomposition to avoid numerical issues that can arise with very small Eigenvalues.

A.2 Log-Determinant

Finally, we can apply the matrix determinant lemma using the definition of $\mathbf{M}$ above to give

$$\log \det \boldsymbol{\Sigma} = \log \det ((\mathbf{LL}^\top)^{-1} + \mathbf{RR}^\top) = -\log \det(\mathbf{LL}^\top) + \log \det(\mathbf{M}), \quad (25)$$

where $\log \det(\mathbf{M})$ falls out of the Eigenvalue decomposition required for the log probability above.

B Segmentation

B.1 Label-wise Accuracy

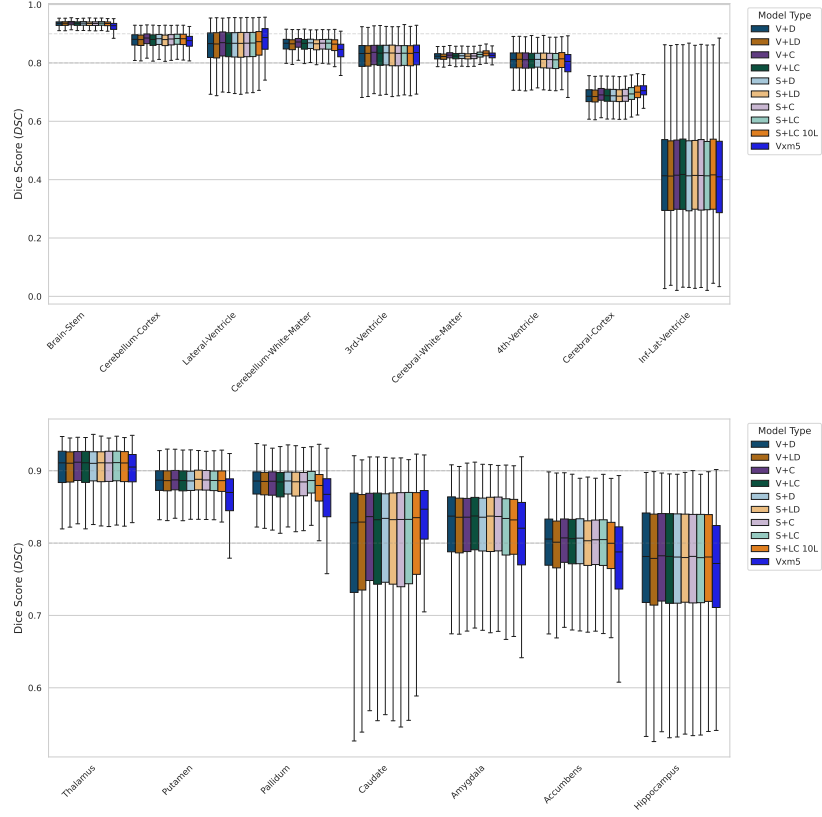


Fig. 6: Boxplots of propagated segmentation accuracy for different labelled region for different model variants, when using the average of sampled fields $\bar{\mathbf{Z}}$.

C Selection Gap

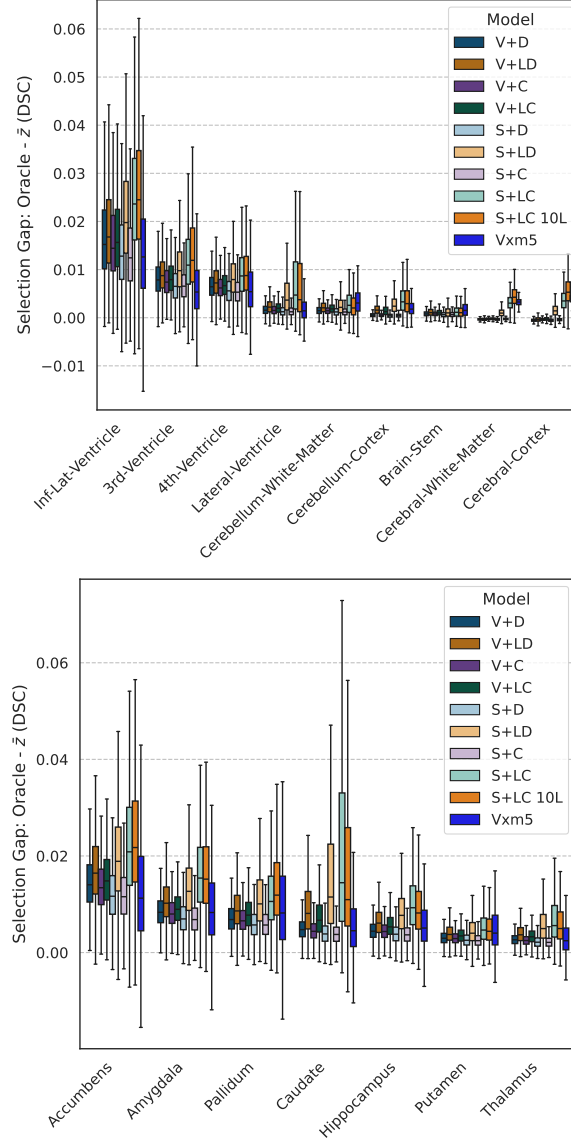


Fig. 7: Boxplot highlighting the quality of the Oracle sample drawn from the approximate posterior distribution for the full set of segmentation regions. Larger Dice Similarity Coefficient (DSC) values indicated greater segmentation accuracy.

D Calibration

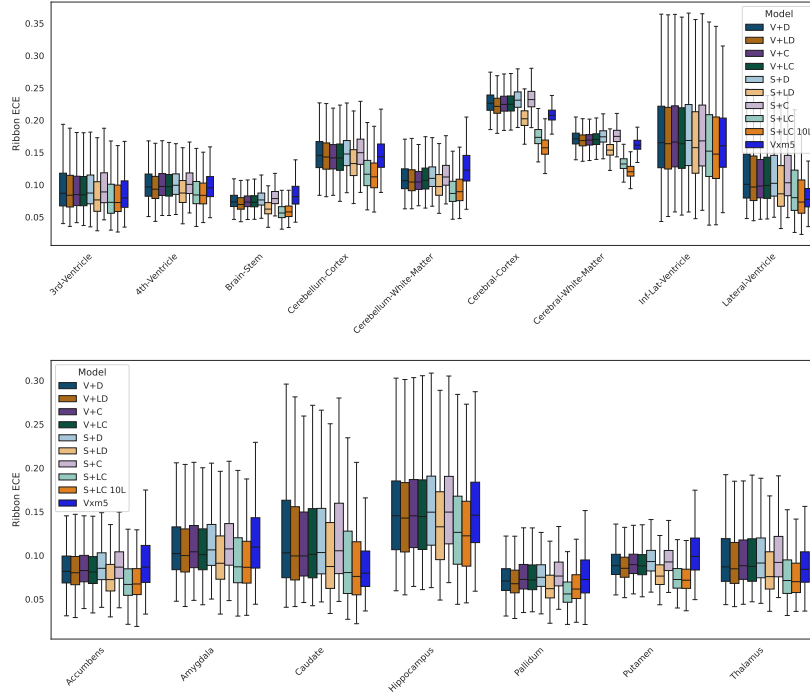


Fig. 8: Expected calibration error (ECE) boxplots across all model variants and structures. ECE was calculated in a region defined by a 3 voxels dilated segmentation mask.

We provide boxplots showing the calibration metrics across all model variance and structures in Fig. 8 and Fig. 9

In Tab. 3 shows a companion representation to Tab. 2 where we instead use displacement entropy for calculating AUSE, and correlating with the Dice score. Here we find that displacement entropy is negatively correlated with Dice for all structures, although the level of entropy is much lower for all except our best calibrated model variant. While displacement entropy quantifies the geometric variance of the deformation field, it captures invisible jitter in homogeneous regions that does not impact anatomical overlap; in contrast, label entropy directly measures the probabilistic risk of misclassification at structural boundaries, providing a more anatomically relevant measure of uncertainty.

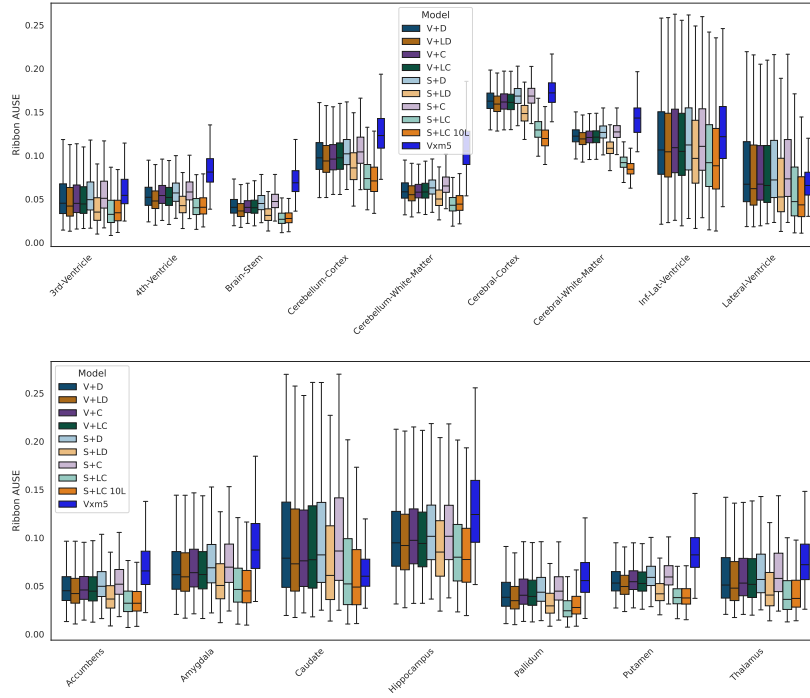


Fig. 9: Area Under Sparse Error (AUSE) boxplots across all model variants and structures. AUSE was calculated in a region defined by a 3 voxels dilated segmentation mask using the label entropy.

Table 3: Uncertainty Calibration Metrics: Area Under Sparse Error (AUSE) based on displacement entropy calculated using a dilated (by 3 voxels) segmentation mask. The last column is the Spearman correlation of label entropy and Dice. All fields show the average and std-dev over structures.

Model	AUSE _{disp} ↓	Displacement Entropy (mm ²)	Spearman r: Entropy & DSC
V+D	0.1351 (0.017)	0.0494 (0.001)	-0.648 (0.121)
V+LD	0.1343 (0.017)	0.0635 (0.002)	-0.430 (0.168)
V+C	0.1312 (0.015)	0.0449 (0.001)	-0.628 (0.138)
V+LC	0.1323 (0.015)	0.0571 (0.005)	-0.530 (0.221)
S+D	0.1366 (0.018)	0.0432 (0.006)	-0.199 (0.059)
S+LD	0.1272 (0.017)	0.1128 (0.014)	-0.549 (0.166)
S+C	0.1339 (0.017)	0.0403 (0.005)	-0.311 (0.099)
S+LC	0.1205 (0.016)	0.1808 (0.037)	-0.595 (0.156)
S+LC 10L	0.1186 (0.016)	0.2270 (0.074)	-0.624 (0.165)
Vxm	0.1223 (0.014)	0.1428 (0.0909)	-0.571 (0.199)

E Training Stability

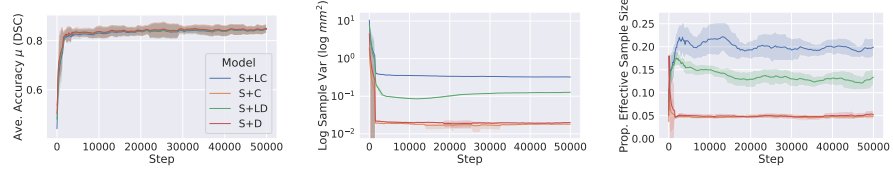


Fig. 10: Average accuracy (left), displacement variance (mid) and proportional effective sample size ($1.0/\sum_i w_i^2$, right) on the val set.

Fig. 10 demonstrates that accuracy, displacement variance and proportional effective sample size (ESS) stabilize smoothly from the cold start, refuting weight degeneracy and confirming that loss signals reliably steer poor early samples.

F Sensitivity Analysis

F.1 Number of low-rank components

Table 4: Registration performance and folding rates across OASIS test pairs modifying the number of low-rank components, with all other hyperparameters held constant. We report the Dice Similarity Coefficient (DSC) for the mean prediction ($\mu_{\mathbf{Z}}$), the mean of resampled fields ($\bar{\mathbf{Z}}$), and the Oracle (the single sample from the posterior with the highest DSC). Fold % represents the percentage of voxels with a non-positive Jacobian determinant. Standard deviations are shown in parentheses. High DSC values and lower fold % are preferable.

Model	μ (DSC)	$\bar{\mathbf{Z}}$ (DSC)	Oracle (DSC)	μ fold %	$\bar{\mathbf{Z}}$ fold %	Oracle fold %
S+LC	0.779 (0.05)	0.804 (0.04)	0.815 (0.03)	0.001 (0.00)	0.453 (0.18)	0.458 (0.36)
5L						
S+LC	0.788 (0.04)	0.802 (0.04)	0.814 (0.03)	0.004 (0.01)	0.240 (0.14)	0.292 (0.25)
10L						
S+LC	0.794 (0.04)	0.801 (0.04)	0.812 (0.03)	0.017 (0.01)	0.143 (0.11)	0.201 (0.18)
20L						
S+LC	0.794 (0.04)	0.801 (0.04)	0.812 (0.03)	0.011 (0.01)	0.115 (0.07)	0.163 (0.13)
25L						
S+LC	0.796 (0.04)	0.800 (0.04)	0.810 (0.04)	0.024 (0.02)	0.092 (0.07)	0.116 (0.10)
30L						

We evaluate the impact of the number of low-rank components in our S+LC approach in Tab. 4 and Tab. 5. We find that 10 components provides an optimal number for uncertainty calibration. However, as all other experiments are conducted with 25 components we report this as our main result.

F.2 Temperature scaling

We provide results for adjusting the Temperature value T in Tab. 6 and Tab. 7. We note that there may be small gains in further optimise the temperature choice.

Table 5: Uncertainty Calibration Metrics comparing the number of low-rank components, with all other hyperparameters held constant. Area Under Sparse Error (AUSE) based on label entropy and Expected Calibration Error (ECE) calculated using a dilated (by 3 voxels) segmentation mask. The last column is the Spearman correlation of label entropy and Dice. All fields show the average and std-dev over structures.

Model	$AUSE_{label} \downarrow$	$ECE \downarrow$	Label Entropy	Mean Spearman r: Entropy & DSC
S+LC 5L	0.0646 (0.018)	0.1030 (0.022)	0.1128 (0.011)	-0.400 (0.224)
S+LC 10L	0.0636 (0.018)	0.1028 (0.023)	0.1156 (0.010)	-0.419 (0.217)
S+LC 20L	0.0681 (0.019)	0.1087 (0.024)	0.1046 (0.007)	-0.376 (0.221)
S+LC 25L	0.0659 (0.019)	0.1064 (0.025)	0.1079 (0.006)	-0.325 (0.218)
S+LC 30L	0.0725 (0.020)	0.1143 (0.026)	0.0940 (0.005)	-0.241 (0.194)

Table 6: Registration performance and folding rates across OASIS test pairs modifying temperature values T , with all other hyperparameters held constant. We report the Dice Similarity Coefficient (DSC) for the mean prediction ($\mu_{\mathbf{Z}}$), the mean of resampled fields ($\bar{\mathbf{Z}}$), and the Oracle (the single sample from the posterior with the highest DSC). Fold % represents the percentage of voxels with a non-positive Jacobian determinant. Standard deviations are shown in parentheses. High DSC values and lower fold % are preferable.

Model	μ (DSC)	$\bar{\mathbf{Z}}$ (DSC)	Oracle (DSC)	μ fold %	$\bar{\mathbf{Z}}$ fold %	Oracle fold %
S+LC $T = 2.0$	0.800 (0.04)	0.801 (0.04)	0.809 (0.04)	0.043 (0.04)	0.073 (0.07)	0.079 (0.08)
S+LC $T = 3.0$	0.794 (0.04)	0.801 (0.04)	0.812 (0.03)	0.011 (0.01)	0.115 (0.07)	0.163 (0.13)
S+LC $T = 4.0$	0.792 (0.04)	0.802 (0.04)	0.813 (0.03)	0.009 (0.01)	0.185 (0.12)	0.224 (0.19)

Table 7: Uncertainty Calibration Metrics across temperature values T , with all under parameters held constant: Area Under Sparse Error (AUSE) based on label entropy and Expected Calibration Error (ECE) calculated using a dilated (by 3 voxels) segmentation mask. The last column is the Spearman correlation of label entropy and Dice. All fields show the average and std-dev over structures.

Model	$AUSE_{label} \downarrow$	$ECE \downarrow$	Label Entropy	Mean Spearman r: Entropy & DSC
S+LC $T = 2.0$	0.0775 (0.020)	0.1207 (0.026)	0.0780 (0.003)	-0.119 (0.146)
S+LC $T = 3.0$	0.0659 (0.019)	0.1064 (0.025)	0.1079 (0.006)	-0.325 (0.218)
S+LC $T = 4.0$	0.0650 (0.018)	0.1044 (0.023)	0.1114 (0.009)	-0.408 (0.219)

G Baseline Model Details

We compare our approach against a deep ensemble of 5 Voxelmorph [9] models. These models were trained with mean-square error, diffusion regularisation, batch size 4 and $\lambda = 0.01$ with equivalent model complexity to ours. Fields were produced at half the image size and upsampled with cubic interpolation, which matches our approach. We also compare against a resolution matched deterministic method, Convex Adam [14] with MIND features [13], and also a full-resolution Voxelmorph ensemble. We present a brief summary of the results in Tab. 8 and note that the full resolution Voxelmorph has unsurprisingly higher accuracy, but still suffers from worse label AUSE (with similar ECE). SIR based models could also be tractably trained and run at full resolution given sufficient computational resource.

Table 8: Comparison of model variants with Convex Adam (CA) with MIND features, a Voxelmorph 5-model ensemble at matched resolution (Vxm5) and a Voxelmorph 5-model ensemble at maximum resolution

	μ (DSC)	$\bar{\mathbf{Z}}$ (DSC)	Oracle (DSC)	$AUSE_{label} \downarrow$	$ECE \downarrow$	Label Entropy
V+D	0.800 (0.04)	0.800 (0.04)	0.805 (0.04)	0.085 (0.02)	0.129 (0.03)	0.063 (0.00)
S+LD	0.798 (0.04)	0.800 (0.04)	0.809 (0.04)	0.074 (0.02)	0.116 (0.03)	0.090 (0.00)
S+LC	0.794 (0.04)	0.801 (0.04)	0.812 (0.03)	0.066 (0.02)	0.107 (0.03)	0.108 (0.01)
S+LC 10L	0.788 (0.04)	0.802 (0.04)	0.814 (0.03)	0.064 (0.02)	0.103 (0.02)	0.115 (0.01)
CA	0.788 (0.04)	0.788 (0.04)	0.788 (0.04)	-	-	-
Vxm5	0.798 (0.03)	0.798 (0.03)	0.804 (0.03)	0.101 (0.02)	0.122 (0.02)	0.070 (0.01)
Vxm5 max	0.818 (0.03)	0.818 (0.03)	0.823 (0.03)	0.081 (0.02)	0.099 (0.02)	0.079 (0.01)

H Uncertainty Metric Details: ECE and AUSE

To evaluate the quality of the predicted spatial uncertainty, we compute the Expected Calibration Error (ECE) and Area Under the Sparsification Error (AUSE) over a localised spatial domain where registration ambiguity is likely to be most prevalent.

H.1 Domain Localization and Probability Mapping

Let Ω denote the volumetric image domain. We define the anatomical ribbon $\Omega_R \subset \Omega$ via a 3-voxel morphological dilation of the ground-truth anatomical label boundaries, containing $N = |\Omega_R|$ evaluation voxels. For each voxel $x \in \Omega_R$, the empirical posterior probability $\hat{p}(x)$ of a given anatomical structure is computed as the average across K warped label samples:

$$\hat{p}(x) = \frac{1}{K} \sum_{k=1}^K Y_k(x) \quad (26)$$

where $Y_k(x) \in \{0, 1\}$ is the propagated binary label for sample k . The binary voxel error map is given by $E(x) = |\mathbb{I}(\hat{p}(x) > 0.5) - Y_{GT}(x)| \in \{0, 1\}$, where $Y_{GT}(x)$ is the ground-truth target label and $\mathbb{I}$ is the indicator function.

H.2 Expected Calibration Error (ECE)

ECE measures the absolute discrepancy between the model's confidence and its actual empirical accuracy. The voxels within Ω_R are partitioned into M equally spaced confidence bins $B_m = (\frac{m-1}{M}, \frac{m}{M}]$ based on $\hat{p}(x)$. The ECE is computed as a weighted average across all bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} \left| \text{conf}(B_m) - \text{acc}(B_m) \right| \quad (27)$$

where the mean confidence $\text{conf}(B_m)$ and empirical accuracy $\text{acc}(B_m)$ of bin B_m are defined as:

$$\text{conf}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} \hat{p}(x), \quad \text{acc}(B_m) = \frac{1}{|B_m|} \sum_{x \in B_m} Y_{GT}(x) \quad (28)$$

H.3 Area Under the Sparsification Error (AUSE)

AUSE assesses the error-ranking capability of an uncertainty metric $u(x)$ (evaluated independently for displacement variance and label entropy). Voxels in Ω_R are permuted via two separate ordering functions:

- **Observed Sorting (π_{unc}):** Ordered by descending uncertainty: $u(\pi_{unc}(1)) \geq u(\pi_{unc}(2)) \geq \dots \geq u(\pi_{unc}(N))$.

- **Oracle Sorting** (π_{err}): Ordered by descending true error: $E(\pi_{err}(1)) \geq E(\pi_{err}(2)) \geq \dots \geq E(\pi_{err}(N))$.

Sparsification curves $c_{obs}(i)$ and $c_{ora}(i)$ track the remaining mean error after sequentially removing the highest $N - i$ entries according to their respective permutations:

$$c_{obs}(i) = \frac{1}{i} \sum_{j=1}^i E(\pi_{unc}(j)), \quad c_{ora}(i) = \frac{1}{i} \sum_{j=1}^i E(\pi_{err}(j)) \quad (29)$$

The AUSE integrates the gap between the two curves across all remaining voxel fractions $i \in \{1, \dots, N\}$:

$$\text{AUSE} = \frac{1}{N} \sum_{i=1}^N (c_{obs}(i) - c_{ora}(i)) \quad (30)$$

An AUSE of 0 denotes perfect error-ranking capability, demonstrating that the model's localised variance directly identifies registration failures.