BioMedInformatics

Article

A Guarded Hybrid Pipeline for Rigid FLAIR–T1 MRI Registration: Learned Initialisation of Iterative Mutual-Information Refinement

Zakaria Said [1,*], Fatima-Ezzahraa Ben-Bouazza [2,3,4] and Mounir Mekkour [1]

[1] Laboratory L2MASI, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed Ben Abdellah University, B.P. 1796, Atlas Fez, Fez 30000, Morocco; mounir.mekkour@usmba.ac.ma
[2] Artificial Intelligence Research and Application Laboratory (AIRA Lab), Hassan I University, Settat 26000, Morocco; fatimaezzahraa.benbouazza@uhp.ac.ma
[3] BRET Lab, Mohammed VI University of Sciences and Health, Casablanca 82403, Morocco
[4] LaMSN—La Maison des Sciences Numériques, 93210 Saint-Denis, France
* Correspondence: zakaria.said@usmba.ac.ma

Abstract

Background: Rigid alignment of multimodal MRI is a prerequisite for neurological post-processing, yet classical iterative optimisers are accurate only when allowed to converge, at tens of seconds per volume, while learning-based predictors are fast but far less precise. This study asks whether a learned predictor can serve as the initialisation of an iterative optimiser so that both speed and accuracy are obtained. **Methods:** A supervised dual-path convolutional network with a cross-modal attention block (AttentionReg) was trained on BraTS 2020 under a subject-level split and a synthetic rigid-transformation protocol. Its prediction initialises a brain-masked Mattes Mutual-Information optimiser, forming a guarded hybrid pipeline in which a refinement is accepted only when it improves the metric. Eight methods were compared under one protocol on a common subset of 990 held-out pairs from 99 subjects, using a landmark-based Target Registration Error expressed in physical millimetres. **Results:** The proposed hybrid attained a median TRE of 0.81 mm over the evaluated pairs (IQR 0.57–1.12; P95 1.67 mm; 0.3% above 3 mm) at 0.79 s per volume, against 0.57 mm at 25.3 s for the fully converged iterative reference. The subject-level paired difference was 0.23 mm (95% CI 0.19–0.28), statistically significant yet equivalent within a 0.50 mm margin, obtained with a 32-fold reduction in computation time. Used alone, the network reached 3.10 mm; iterative refinement improved it by a mean subject-level paired reduction of 2.49 mm (95% CI 2.32–2.67). An ablation showed the attention block does not change median accuracy relative to a matched network without it. **Conclusions:** A learned initialiser converts an accurate but slow iterative method into an accurate and fast one, providing a proof of concept for real-time rigid multimodal MRI registration.

Keywords: deep learning; rigid registration; cross-modal attention; multimodal MRI; FLAIR; BraTS; Target Registration Error; neurological imaging

Academic Editors: Saurav Mallik, Priyanka Roy and Zhongming Zhao

Received: 28 June 2026
Revised: 7 September 2026
Accepted: 17 September 2026
Published: 21 September 2026

1. Introduction

Medical imaging holds a distinctive place in modern clinical practice, enabling physicians to study internal anatomy without invasive procedures and to ground their diagnoses in objective, reproducible evidence, whether identifying haemorrhages, delineating lesions, or detecting aberrant cellular formations. Recent advances in machine learning have

opened new avenues for assisting clinicians by automating the most time-consuming and repetitive steps of image interpretation [1,2].

This work investigates the potential of deep learning for MRI registration, a process that underpins diagnostic support tools for patients with brain pathologies including tumours, neurodegenerative illnesses, and cerebrovascular accidents. Patient movement during MRI acquisition frequently produces misalignment across sequences. When this movement is confined to rotations and translations, which is the typical situation in brain imaging, the resulting displacement is termed rigid misalignment. This offset can profoundly affect radiological interpretation and compromise any downstream processing that depends on accurate spatial correspondence. Precise alignment is a prerequisite for segmenting anatomical structures, tracking anomalous changes over time, evaluating post-treatment response, and detecting early relapse.

Registration algorithms compensate for these misalignments. The conventional approach relies on iterative optimisation, progressively adjusting image positions to minimise a cost function. Although iterative methods are accurate when allowed to converge, their execution times reach tens of seconds per volume (in our own measurements, up to 25 s for a fully converged multimodal rigid solve, Section 3.7), making them poorly suited for time-sensitive clinical use unless suitably initialised.

1.1. The Formalism of Rigid Image Registration

Rigid image registration aligns two acquisitions while preserving their geometric integrity, a procedure indispensable across diagnosis, treatment planning, and disease monitoring [3]. Formally, it seeks the optimal transformation T maximising a similarity measure $S(I, J, T)$ [4]:

$$\hat{T} = \arg\max_{T} S(I, J, T). \quad (1)$$

Because rigid transformations exclude deformation, the parameter space is compact: three parameters in 2D and six (three translations, three rotations) in 3D [5]. Conventional strategies such as Gradient Descent [6] and Powell's Method [7] carry two fundamental limitations: susceptibility to local optima in multimodal settings, and prohibitive computational cost for real-time use.

1.2. Deep Learning Approaches to Rigid Registration

Convolutional Neural Networks (CNNs) trained on paired images learn to predict similarity scores that guide transformation estimates [4,8]. End-to-end architectures bypass optimisation entirely:

- Supervised registration regresses rotation and translation parameters directly from image pairs [9].
- Unsupervised registration optimises via a learned similarity loss without ground-truth labels [10].
- Self-supervised registration uses cycle-consistency or reconstruction losses as supervision signals [11].
- Reinforcement learning frames registration as sequential decision-making with reward signals [12].

Transformer-based architectures have further extended the field. TransMorph [13] introduced a Swin Transformer backbone for deformable registration; ViT-V-Net [14] combined vision transformers with a V-Net decoder for unsupervised volumetric registration. Their use of attention mechanisms to establish spatial correspondences directly informs the cross-modal attention design adopted here for the rigid multimodal setting.

1.3. Clinical Motivation and Contribution

The choice of FLAIR and T1-weighted sequences is clinically motivated. In stroke imaging, FLAIR reveals hyperintense regions indicative of acute infarction while T1 provides a detailed assessment of tissue damage [15]. In epilepsy, FLAIR highlights cortical dysplasias and hippocampal sclerosis, whereas T1 supplies anatomical detail for pre-surgical evaluation [15]. In Alzheimer's disease, T1 images trace brain atrophy while FLAIR exposes white matter alterations [16]. In multiple sclerosis, FLAIR maps demyelinating plaques while T1 identifies hypointense lesions associated with severe tissue damage [17].

Rather than positioning the learned model as a replacement for classical iterative tools, this work evaluates both families under an identical protocol and, crucially, combines them. Standard intensity-based metrics applied to FLAIR-to-T1 registration produce noisy optimisation landscapes because of the strong contrast inversion between modalities, so iterative optimisers such as ANTs [18] and SimpleITK's rigid Mutual-Information optimiser are accurate only when allowed to converge fully, at a cost of tens of seconds per volume, and are otherwise prone to local minima. The learned model delivers a modality-agnostic estimate in a single forward pass; we show that using this estimate to initialise the iterative optimiser resolves the trade-off, yielding accuracy within a third of a millimetre of the converged reference in under one second.

This paper makes the following contributions:

- A guarded hybrid registration pipeline in which a learned predictor initialises an iterative Mutual-Information optimiser. On a common subset of 990 held-out pairs, it attains a median Target Registration Error of 0.81 mm at 0.79 s per volume, against 0.57 mm at 25.3 s for the fully converged iterative reference: a difference of 0.23 mm, equivalent within a 0.50 mm margin, at a 32-fold reduction in computation time.
- A demonstration that the learned component is valuable as an *initialiser* rather than as a standalone registration method. Used alone, the network reaches 3.10 mm; the refinement stage reduces this by 2.49 mm (95% CI 2.32–2.67), a near-complete separation of the two distributions.
- A controlled evaluation of eight methods under a single protocol and a single error measure expressed in physical millimetres, with subject-level statistics, bootstrap confidence intervals, Holm-corrected paired tests and equivalence testing.
- An ablation isolating the cross-modal attention block, reported honestly: it does not change median accuracy relative to a matched network without attention.

2. Materials and Methods

2.1. Datasets

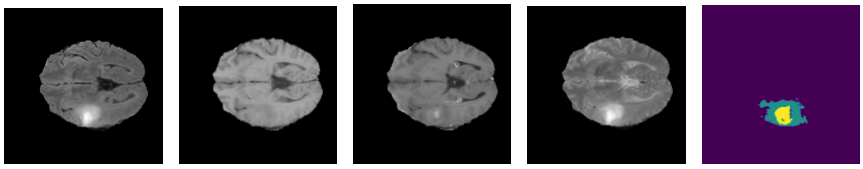
BraTS 2020 Dataset

The BraTS 2020 dataset [19–21] is a standardised collection of multimodal brain tumour MRI scans preprocessed according to the MICCAI challenge framework: co-registration to a common anatomical template, interpolation to 1 mm^3 isotropic resolution, and skull-stripping. The dataset comprises four sequences: T1-weighted, post-contrast T1-weighted (T1Gd), T2-weighted, and FLAIR, from diverse institutions (Figure 1).

For the registration task, FLAIR is designated as the fixed volume and T1 as the moving volume. Intensity values are normalised to $[0, 1]$ and volumes are resized to $96 \times 96 \times 32$ voxels. Because BraTS volumes are pre-aligned, rigid misalignments are introduced synthetically (Section 2.2).

Subject-level partitioning. To preclude data leakage, the dataset is partitioned by subject rather than by transformed sample. Of the 494 available subjects, 99 are held out for testing and the remaining 395 form the development pool, which is split at the patient level into 356 training and 39 validation subjects (90%/10%). With 30 synthetic

transformations per subject this yields 10,680 training pairs, 1170 validation pairs and 2970 test pairs. Because all transformations of a given subject are confined to a single fold, no subject and no anatomy are shared across partitions.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 1. The four MRI modalities of the BraTS 2020 dataset with corresponding ground-truth segmentation. (**a**) FLAIR. (**b**) T1-weighted. (**c**) T1 contrast-enhanced. (**d**) T2-weighted. (**e**) Ground-truth segmentation mask.

2.2. Rigid Transformation Formalism

2.2.1. Mathematical Formulation

A rigid transformation maps each point $p = [x, y, z]^T$ to $p' = [x', y', z']^T$ through a 4×4 homogeneous matrix combining a rotation and a translation, so that the six degrees of freedom are three translations (q_x, q_y, q_z) and three rotation angles $(\theta_x, \theta_y, \theta_z)$ about the coordinate axes. Because the composition of rotations is not commutative, the order in which they are applied must be stated explicitly for the results to be reproducible. Throughout this work the rotations are applied in the order x, y, z and the translation last:

$$M_{\text{rigid}} = M_{\text{trans}} \cdot M_z(\theta_z) \cdot M_y(\theta_y) \cdot M_x(\theta_x). \tag{2}$$

2.2.2. Implementation via SimpleITK

Transformations are applied using SimpleITK [22] through resampling with linear interpolation. SimpleITK's global domain transform is:

$$T(p) = A(p - c) + t + c, \quad \text{Offset} = t + c - Ac, \tag{3}$$

where A is the rotation matrix, c the centre of rotation, and t the translation vector. Since SimpleITK maps from output to input space, the inverse of the predicted transformation is applied to the moving volume.

2.3. Image Geometry and Units

Because the principal results are expressed in millimetres, the relationship between the voxel grid used by the network and physical space is stated explicitly.

BraTS volumes are distributed at $155 \times 240 \times 240$ voxels with 1 mm isotropic spacing. For training and inference, they are resampled by trilinear interpolation to $32 \times 96 \times 96$ voxels, preserving the field of view. The physical voxel size after resampling is therefore anisotropic: $2.5 \times 2.5 \times 4.84$ mm, the larger value corresponding to the through-plane axis, which is reduced from 155 to 32 samples.

Synthetic transformations are generated in the resampled grid: translations are expressed in resampled voxels (bounded at ± 10 per axis) and rotations in degrees (bounded at ± 15 per axis). Target Registration Error and warping index are computed on landmark sets carried in the physical coordinate system of the resampled volume, that is, with the anisotropic spacing above applied to the displacement components before the Euclidean norm is taken. All error values reported in this work are therefore physical millimetres. For continuity with the resampled grid, and so that the conversion can be verified, the corresponding values in resampled-voxel units are also given in the complete comparison

of Section 3.7; across methods, the ratio between the two lies between 2.94 and 4.39, with a mean of 3.29.

Two consequences follow and are stated plainly. First, a transformation that is rigid in the resampled grid corresponds, in physical space, to a rotation conjugated by an anisotropic scaling, and is therefore not exactly a rigid physical motion except for pure translations; the synthetic protocol should be read accordingly. Second, the 3 mm reference threshold used in the earlier literature is smaller than a single voxel along the through-plane axis, so it is a demanding criterion for any method operating at this resolution; success rates are consequently reported across a range of thresholds (Section 3.3.3) rather than at a single cut-off.

2.4. Model Architecture

The proposed architecture is adapted from the cross-modal attention model of Song et al. [23], originally designed for MRI-to-transrectal ultrasound registration. Two targeted modifications suit the FLAIR/T1 brain context: (1) reduced depth of the feature extraction branches and (2) fully re-tuned hyperparameters. The architecture is illustrated in Figure 2 and consists of three functional modules.

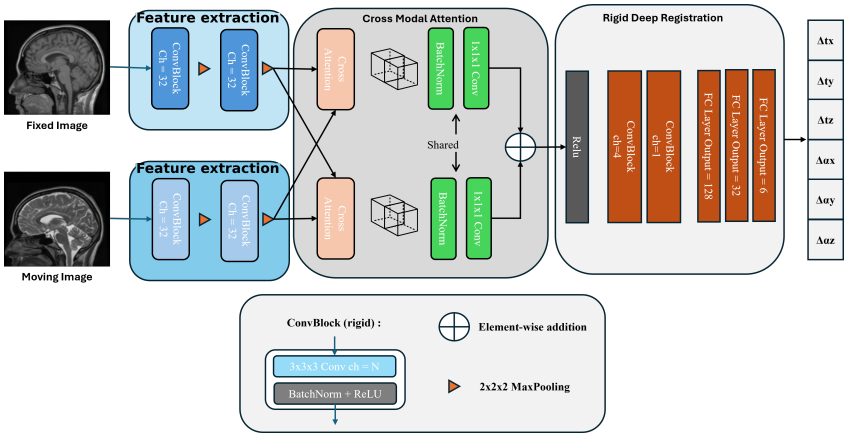


Figure 2. Overview of the proposed network architecture, adapted from [23] with reduced feature extractor depth and re-tuned hyperparameters for the FLAIR/T1 brain registration context. The three modules are: (1) parallel feature extraction branches; (2) cross-modal attention block; and (3) fully connected regression head predicting the six rigid transformation parameters.

2.4.1. Feature Extraction

Two independent feature extractors, one per modality, apply convolutional layers, batch normalisation, ReLU activations, and down-sampling to produce hierarchical, semantically rich representations abstracting away modality-specific intensity patterns.

2.4.2. Cross-Modal Attention

Given a primary feature map P and a cross-modal feature map C (dimensions $C \times L \times W \times H$), attention weights are:

$$a_{ij} = \frac{\exp\left(\theta(c_i)^\top \phi(p_j)\right)}{\sum_j \exp\left(\theta(c_i)^\top \phi(p_j)\right)}, \tag{4}$$

where $\theta(\cdot)$ and $\phi(\cdot)$ are linear projections into a shared embedding space. The attention-weighted output $y_i = \sum_j a_{ij} g(p_j)$ is concatenated with C:

$$Z = [Y, C]. \tag{5}$$

2.4.3. Rigid Transformation Prediction

Fully connected layers regress the six rigid parameters (three rotations, three translations) from the fused representation Z.

Detailed configuration. Each of the two feature-extraction branches receives a single-channel volume ($1 \times 32 \times 96 \times 96$) and applies two convolutional blocks (Conv3D $3 \times 3 \times 3$, 32 filters, stride 1, padding 1; BatchNorm3D; ReLU), each followed by max-pooling ($3 \times 3 \times 3$, stride 2 then stride $(1, 2, 2)$), reducing the volume to $32 \times 8 \times 24 \times 24$. The cross-modal attention block is a non-local operation [24] in embedded-Gaussian mode: query, key, and value projections (θ, ϕ, g) are $1 \times 1 \times 1$ convolutions reducing 32 channels to an internal dimension of 16; the affinity map is normalised by softmax; and the output projection W_z is a $1 \times 1 \times 1$ convolution back to 32 channels, zero-initialised so that the block starts as an identity mapping. Each branch is refined by the other branch's features, and the two are concatenated (64 channels). The regression head reduces this via three convolutions ($1 \times 1 \times 1$: 64→16; $3 \times 3 \times 3$: 16→4; $3 \times 3 \times 3$: 4→1) and three fully connected layers ($9216 \rightarrow 128 \rightarrow 32 \rightarrow 6$). The ablation model (FeatureReg) is byte-identical except that the two attention calls are removed and the branch features are concatenated directly; it differs from the proposed model by only 4384 parameters (Table 1), isolating the effect of the attention block. Layer-wise details and complexity are given in Table 1.

Table 1. Architecture and complexity of the evaluated learning-based models. The attention block adds 4384 parameters (0.35%) over the ablation, isolating its effect. Latency is the mean single-volume network forward pass measured in isolation on the evaluation hardware. The per-volume time reported in Section 3.7 (0.108 s) is larger because it covers the complete registration operation, including volume loading and resampling.

Model	Attention	Parameters	MFLOPs	Latency (ms)
VoxelMorph-rigid	—	163,014	1335	80.0
FeatureReg (ablation)	no	1,244,393	2663	80.5
AttentionReg (proposed)	cross-modal non-local	1,248,777	2703	83.8

2.5. Training Pipeline

All models are trained under identical conditions to ensure a controlled comparison. Weights are initialised with Xavier uniform initialisation; optimisation uses stochastic gradient descent without momentum, with L2 weight decay (10^{-2}) and a multi-step learning-rate schedule, with a fixed random seed governing initialisation and batch ordering. The ablation (FeatureReg) is trained with the same seed, schedule, data, and number of epochs as the proposed model, so that any performance difference is attributable to the attention block alone.

The loss is the mean squared error (MSE) computed jointly over the six transformation parameters. Because rotations (degrees) and translations (resampled voxels) occupy comparable numerical ranges under the chosen transformation bounds ($\pm 15°$, ± 10 voxels), the unweighted joint MSE does not implicitly favour one group; the per-parameter error analysis in Section 3 confirms that both families are learned, with the residual error concentrated on rotation, an asymmetry the hybrid pipeline (Section 2.8) subsequently removes. The learning-rate schedule was defined with milestones at epochs 15, 120, 170 and 250 for a longer training budget than the one ultimately used; within the 50 epochs actually run, only the first milestone is reached, so exactly one reduction in the learning rate is applied, at epoch 15. This is visible as the single step in the loss curves reported in Section 3.2. Key hyperparameters are summarised in Table 2.

Table 2. Training hyperparameters.

Component	Value
Optimizer	SGD, no momentum, weight decay 10^{-2}
Initial learning rate	0.001
LR scheduler	MultiStepLR, milestones 15/120/170/250, $\gamma = 0.3$
Random seed	42 (fixed for all models)
Regularisation	L2 weight decay
Loss function	Mean Squared Error (MSE)
Weight initialisation	Xavier uniform
Batch size	8
Training epochs	50
Train/validation split	90%/10%, subject-level
Input volume dimensions	$96 \times 96 \times 32$ voxels
Augmentations per volume pair	30 synthetic rigid transforms
Max translation	± 10 resampled voxels per axis
Max rotation	$\pm 15^{\circ}$ per axis

2.6. Evaluation Metrics

Three complementary metrics are used.

Absolute Error (AE) measures per-parameter prediction accuracy:

$$AE = |p - g|, \tag{6}$$

where p is the predicted value and g its ground-truth counterpart.

Mutual Information (MI) quantifies shared information between the fixed and registered images:

$$MI(X, Y) = \sum_{i=1}^{n} \sum_{j=1}^{m} P_{XY}(i, j) \log\left(\frac{P_{XY}(i, j)}{P_X(i)\, P_Y(j) + \epsilon} \right). \tag{7}$$

Target Registration Error (TRE) measures residual spatial misalignment via Euclidean distance between corresponding anatomical landmarks in the fixed and registered volumes:

$$TRE = \sqrt{(x_f - x_r)^2 + (y_f - y_r)^2 + (z_f - z_r)^2}. \tag{8}$$

For each test volume, a fixed set of anatomical landmarks is obtained automatically by detecting salient keypoints on the fixed volume; the corresponding points on the registered volume are located by applying the predicted transformation. Because the ground-truth transformation is known by construction, landmark correspondence is exact and requires no manual annotation, which removes inter-rater variability but means the reported TRE reflects the accuracy of the estimated transformation rather than annotation quality. To complement the landmark-based TRE with a metric that is entirely independent of keypoint detection, we additionally report a *warping index*: the mean residual displacement, in millimetres, between the ground-truth and predicted transformations evaluated over all brain voxels. The warping index is a dense, fully reproducible geometric error measure and is used as the primary accuracy metric for the iterative and hybrid methods, for which it is directly comparable to TRE.

2.7. Baseline Methods

To position the proposed model within both families of registration methods, five baselines are evaluated under the same protocol. Learning-based baselines are FeatureReg, the attention ablation described above, and VoxelMorph-rigid, a compact convolutional regressor adapted from the VoxelMorph encoder [10] to output six rigid parameters and trained identically.

Classical iterative baselines. An iterative optimiser exposes an explicit budget that trades computation time against accuracy, so a single configuration cannot characterise this family. We therefore evaluate two points on that trade-off. SimpleITK-MI-converged is allowed to run essentially unconstrained and establishes the accuracy ceiling attainable by intensity-based rigid registration on these data: regular-step gradient descent, learning rate 2.0, up to 500 iterations, 50% metric sampling, 32 histogram bins, minimum step 10^{-5} and gradient-magnitude tolerance 10^{-8}. SimpleITK-MI-fast represents a speed-constrained deployment in which the budget is capped: gradient descent, learning rate 1.0, 200 iterations, 20% sampling, 50 histogram bins, convergence window 20 with minimum value 10^{-9}. Both use a Mattes Mutual-Information metric over the whole volume with random sampling (fixed seed), linear interpolation, a three-level multi-resolution pyramid (shrink factors 4/2/1, smoothing sigmas 2/1/0 in physical units), optimiser scales estimated from physical shift, and an Euler3D transform centred on the volume centre. ANTs-rigid applies the Advanced Normalization Tools rigid stage with a Mattes metric at default settings [18]. All three are initialised from the identity transform, and per-volume timing is recorded single-threaded on the evaluation hardware (Intel Xeon, 2.00 GHz).

Evaluation subsets. The learning-based models, SimpleITK-MI-fast, and Hybrid-fast are evaluated on all 2970 test pairs. SimpleITK-MI-converged requires approximately 25 s per volume, which would amount to more than thirteen hours of single-threaded computation for the full set; it is therefore evaluated on a stratified subset that retains *all* 99 held-out subjects with 10 transformations each, giving 990 pairs.

ANTs-rigid and Hybrid-converged are evaluated on the same 990-pair subset, so that the three are directly comparable to one another and every subject remains represented. The sample size for each method is stated in Section 3.7. Two denominators therefore appear in this article and are always named: analyses concerning only methods evaluated on every pair, such as the per-parameter error of the initialiser in Table 3, use the full 2970 pairs, whereas every comparison between methods uses the common 990-pair subset.

A note on metric masking. The iterative baselines compute Mutual Information over the entire volume, whereas the hybrid pipeline described below restricts the metric to a brain mask. This difference is deliberate and reflects the two regimes rather than a tuning advantage. Starting from the identity transform, the metric is dominated by gross anatomical overlap and the null background of a skull-stripped volume is immaterial; masking changes little, and the converged baseline indeed reaches 0.57 mm without it. Starting from an initialisation already close to the optimum, however, that same null background flattens the metric locally and allows the optimiser to drift away from a good starting point. Brain masking, together with regular rather than random sampling, is therefore a requirement of local refinement specifically, and we report it as a property of the hybrid configuration rather than of the metric itself.

2.8. Guarded Hybrid Pipeline Design

The central contribution combines the learned predictor with iterative refinement. For each pair, AttentionReg produces a rigid estimate in a single forward pass; this estimate initialises a SimpleITK rigid optimiser with a Mattes Mutual-Information metric masked to the brain region, so that background voxels, which dominate a skull-stripped field and flatten the metric near the optimum, do not corrupt the refinement. Two variants are evaluated, differing only in the optimiser budget applied after initialisation (Hybrid-fast and Hybrid-converged). A guard mechanism accepts the refined transformation only if it improves the Mutual-Information value relative to the network estimate, and otherwise retains the network output. It should be stated precisely what this guarantees: the guard rejects any refinement that fails to improve the chosen Mutual-Information objective, but

because that objective is an intensity-based proxy for geometric alignment, acceptance does not by itself guarantee a lower registration error. We quantify this directly in Section 3.8. The reported hybrid latency is the sum of the network forward pass and the iterative refinement, measured on the same hardware.

Table 3. Summary statistics of per-parameter absolute error across the 2970 test pairs for the proposed initialiser. Translations are converted to physical millimetres using the anisotropic voxel size of Section 2.3; rotations are in degrees. Note that TZ carries the largest physical translation error despite having the smallest error in resampled-voxel units, because the through-plane axis has the coarsest spacing.

Stat.	TX (mm)	TY (mm)	TZ (mm)	RX (°)	RY (°)	RZ (°)
Mean	1.00	0.81	1.42	0.68	0.64	1.05
Std	0.82	0.71	1.19	0.58	0.60	0.93
25%	0.38	0.30	0.52	0.26	0.23	0.38
Median	0.80	0.63	1.15	0.54	0.49	0.84
75%	1.41	1.13	1.99	0.95	0.87	1.47
Max	6.86	7.28	10.17	5.55	6.72	12.01

Figure 3 summarises the complete experimental design: panel (a) contrasts the three learning-based models (the proposed network, the ablation obtained by deleting the attention block, and the compact single-path baseline), while panel (b) shows how the learned prediction is embedded in the guarded hybrid pipeline, together with the two optimiser budgets evaluated.

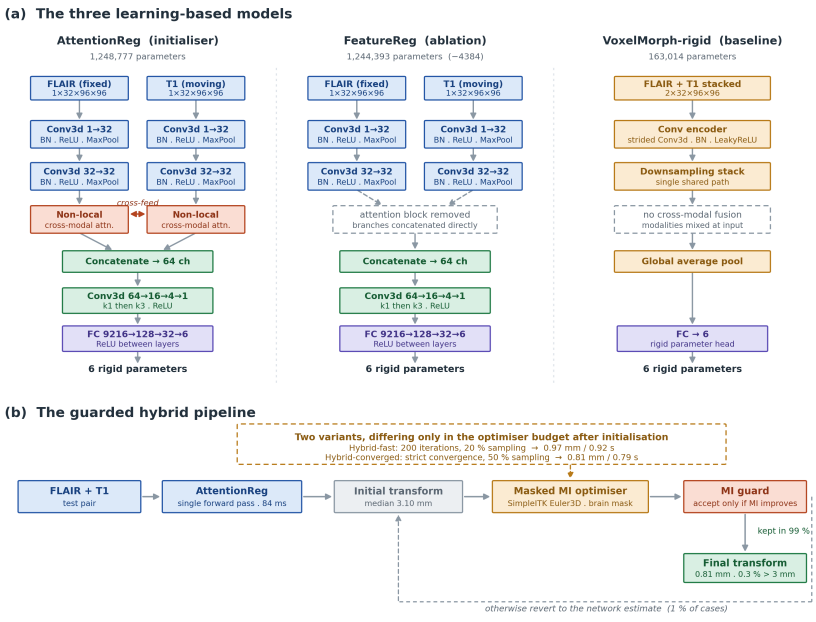


Figure 3. Overview of the evaluated systems. (**a**) The three learning-based models. AttentionReg uses two parallel feature-extraction branches coupled by a cross-modal non-local attention block; FeatureReg is byte-identical except that the attention block is removed and the branches are concatenated directly (a difference of 4384 parameters), isolating the effect of attention; VoxelMorph-rigid stacks both modalities at the input and processes them through a single, far smaller encoder. All three regress the same six rigid parameters. (**b**) The guarded hybrid pipeline: the network prediction initialises a brain-masked Mutual-Information optimiser, and a guard accepts the refinement only when it improves the metric, otherwise reverting to the network estimate. The two variants differ solely in the optimiser budget applied after initialisation.

3. Results

3.1. Frame-Limit Study: Sizing the Operating Range

Before fixing the evaluation protocol, a preliminary scoping study was conducted to determine the displacement range over which the task remains well matched to the model capacity. Three training runs were performed under identical settings, differing only in the maximum translation bound: 10, 20, and 30 resampled voxels per axis, with rotations fixed at $\pm 15^{\circ}$. The intent was not to establish the largest range the network can nominally handle, but to identify the range within which the problem stays well conditioned, so that the study is dimensioned to a task the models solve reliably rather than to one that exceeds their representational capacity.

The outcome is unambiguous (Figure 4). The best validation loss degrades sharply as the range widens: 0.65 at 10 voxels, 2.84 at 20 voxels (a factor of 4.4), and 22.06 at 30 voxels (a factor of 33.8). More telling than the absolute values is the behaviour of the generalisation gap. In the 10-pixel regime, the ratio of validation to training loss stabilises near 3, whereas at 20 and 30 voxels it grows steadily throughout training, reaching 16.6 and 6.7, respectively, by the final epoch, with the training loss continuing to fall while the validation loss stagnates or, in the 30-voxel run, deteriorates after epoch 22. This is the signature of a task whose complexity has outgrown the capacity allocated at the chosen input resolution: the network increasingly fits the training transformations without learning a displacement estimator that transfers to unseen ones.

The 10-voxel range, which also covers the head motion typically encountered in brain MRI under standard protocols, was therefore adopted for all subsequent experiments. This is a deliberate design decision rather than a limitation discovered after the fact: the operating envelope is set where the models are dependable, and extending it is treated as a separate scaling question (Section 4).

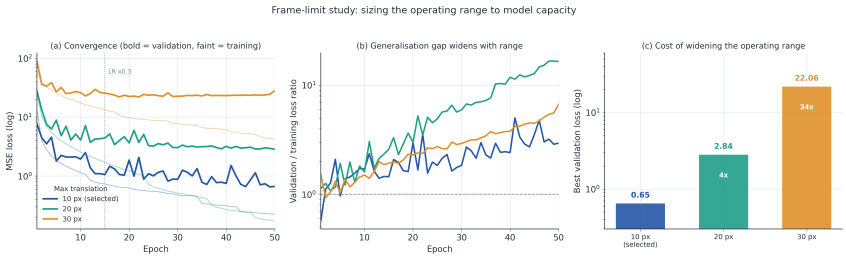


Figure 4. Frame-limit study across three displacement regimes. (**a**) Validation loss (bold) and training loss (faint) on a logarithmic scale; the dotted line marks the learning-rate reduction at epoch 15. (**b**) Ratio of validation to training loss: the generalisation gap widens as the operating range grows, indicating that the network fits the training transformations without generalising to unseen ones. (**c**) Best validation loss per regime, showing the cost of widening the range. The 10-pixel regime was selected for all subsequent experiments.

3.2. Training Convergence

All three learning-based models were then trained within this operating range under an identical protocol (50 epochs, same seed, schedule, and data). Figure 5 reports their training dynamics. The two dual-path models converge to a markedly lower validation loss (best 0.659 for AttentionReg and 0.657 for FeatureReg, at epochs 48 and 49) than the compact VoxelMorph-rigid baseline (best 1.157 at epoch 47), consistent with the accuracy ordering observed at test time. The learning-rate reduction at epoch 15 is clearly visible as a step change in both loss curves. AttentionReg and FeatureReg track one another closely throughout training and converge to within 0.002 of one another, which anticipates the ablation result reported in Section 3.7: the attention block does not alter global convergence.

The decomposition of the validation loss into its rotational and translational components is more informative. For all three models, the rotational term remains roughly six times larger than the translational term at convergence (6.0× for AttentionReg, 6.8× for FeatureReg, 6.0× for VoxelMorph-rigid), and this separation persists across the entire training trajectory. Rotation is therefore the intrinsically harder sub-problem, independently of architecture. This observation from the optimisation dynamics corroborates the per-parameter test set analysis (Section 3.3.1) and motivates the iterative refinement stage of the hybrid pipeline.

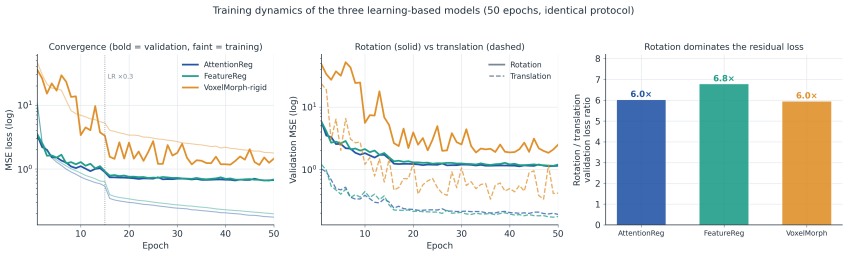


Figure 5. Training dynamics of the three learning-based models over 50 epochs under an identical protocol. (**Left**) Validation loss (bold) and training loss (faint) on a logarithmic scale; the dotted line marks the learning-rate reduction at epoch 15. (**Centre**) Validation loss decomposed into its rotational (solid) and translational (dashed) components. (**Right**) Ratio of rotational to translational validation loss at convergence. Rotation remains approximately six times harder than translation for every architecture, independently of capacity.

3.3. BraTS Benchmark Evaluation

The evaluation was conducted on the 99 held-out test subjects with 30 synthetic transformations each, yielding 2970 test pairs. Three dimensions are assessed: parameter-level prediction accuracy, image-level alignment quality, and spatial registration error.

3.3.1. Parameter-Level Accuracy

Figure 6 compares true and predicted transformation parameters across all six degrees of freedom. Predicted values align closely with the identity diagonal for both translations and rotations, with no systematic over- or underestimation.

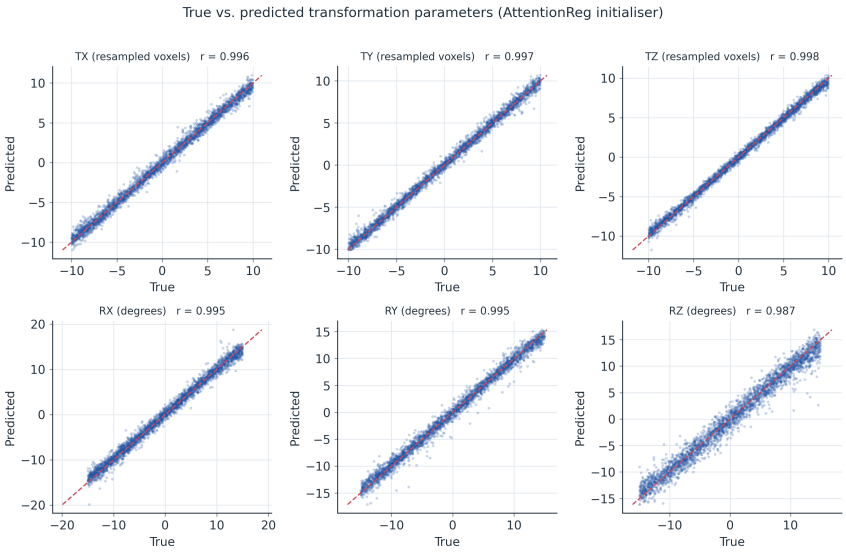


Figure 6. True vs. predicted transformation parameters for the proposed initialiser. Top row: translations TX, TY, TZ in resampled voxels, the units in which the synthetic transformations are generated. Bottom row: rotations RX, RY, RZ in degrees. Points closely following the identity diagonal indicate accurate prediction.

Table 3 and Figure 7 summarise the absolute error distributions. Converted to physical units, the mean translation errors are 1.00, 0.81 and 1.42 mm for TX, TY and TZ. The ordering is instructive: TZ has the smallest error in resampled voxels but the largest in millimetres, because the through-plane axis has the coarsest spacing. Median rotational errors are 0.54, 0.49 and 0.84° for RX, RY and RZ.

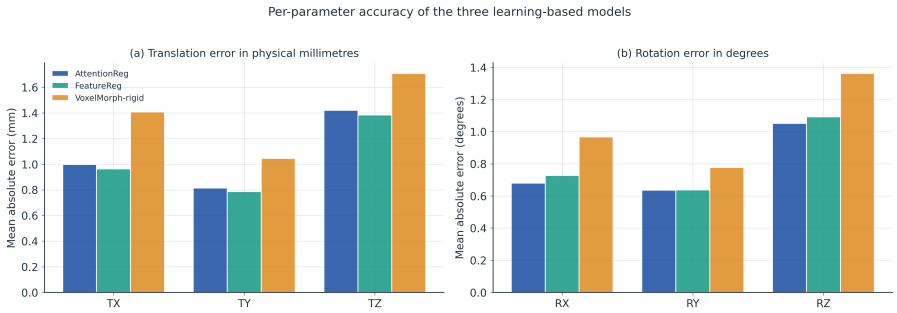
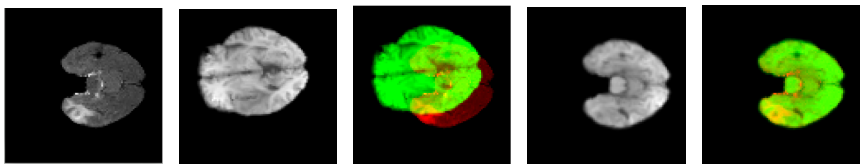


Figure 7. Per-parameter accuracy of the three learning-based models across the test set. (**a**) Translation error converted to physical millimetres using the anisotropic voxel size of Section 2.3; the through-plane axis TZ carries the largest physical error despite the smallest error in voxel units. (**b**) Rotation error in degrees, where the axial rotation RZ is the hardest axis for every model.

3.3.2. Image-Level Alignment Quality

Figure 8 shows the registration process applied to a representative test volume. The pre-registration overlay shows clear spatial mismatch; the post-registration overlay demonstrates tight alignment.



(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 8. Visual representation of the registration process. (**a**) Fixed volume (FLAIR) central slice. (**b**) Displaced moving volume (T1). (**c**) Overlay before registration: clear mismatch visible. (**d**) Registered volume. (**e**) Overlay after registration: tight alignment achieved.

Mutual Information was recomputed from scratch for every stage of the pipeline on the full held-out test set, using a 64-bin joint histogram restricted to brain voxels; the mask is derived from the fixed volume and is therefore identical before and after alignment, so the comparison is not confounded by a changing support. Two reference states bracket the results: the misaligned input, and the volume resampled with the ground-truth transformation, which represents the maximum attainable Mutual Information under this resampling scheme. Figure 9 reports the outcome.

Registration raises the mean Mutual Information from 0.1202 bits in the misaligned state to 0.3296 bits for AttentionReg, a 2.7-fold increase, while the ground-truth reference reaches 0.3992 bits. Four decimals are given so that the fractions below can be verified directly; rounding to three would not reproduce them. Expressing each method as a fraction of the attainable gain is more informative than the raw increase: AttentionReg recovers 75.1% of the available improvement and FeatureReg 75.1%, VoxelMorph-rigid 65.0%, and the guarded hybrid 97.6% (fast) and 99.6% (converged). The hybrid therefore recovers almost all of the attainable gain on an intensity-based criterion, which is consistent with its 0.71 mm geometric residual and confirms that the remaining error is below the level at which Mutual Information can discriminate. The same ordering appears in the geometric residual, which falls from a median of 34.9 mm before registration to 2.42 mm after the

network and 0.71 mm after hybrid refinement. As expected from the ablation, AttentionReg and FeatureReg differ by less than 0.0002 bits on this metric as well (0.3296 vs. 0.3297).

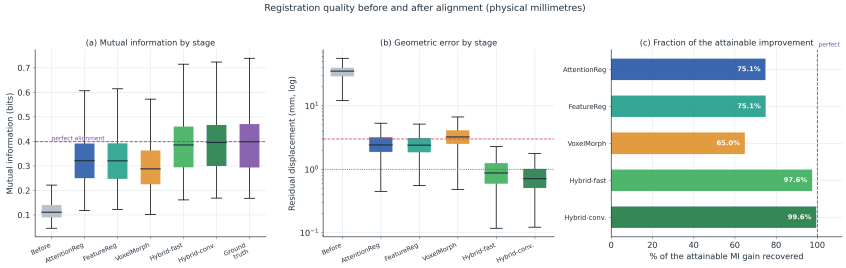


Figure 9. Registration quality before and after alignment on the held-out test set. (**a**) Mutual information at each stage; the dashed line marks the value obtained with the ground-truth transformation, that is, the maximum attainable under this resampling scheme. (**b**) Residual displacement at each stage on a logarithmic axis, from the applied misalignment (median 34.9 mm) down to 0.71 mm after hybrid refinement. All displacements are in physical millimetres. (**c**) Fraction of the attainable Mutual-Information gain recovered by each method. Statistics are computed over 2970 test pairs, except Hybrid-converged, which was run on 990.

3.3.3. Spatial Registration Accuracy

Table 4 and Figure 10 summarise the TRE distribution using subject-level statistics. The proposed pipeline attains a subject-level median TRE of 0.86 mm (95% CI 0.79–0.97 mm), and every subject falls below the 3 mm reference. Used alone, the network reaches 3.18 mm and 63% of subjects exceed that reference, which is why the refinement stage is required rather than optional. The residual error is concentrated on rotation and, within rotation, on the axial axis (RZ), which is the hardest parameter for all three learning-based models (Table 3, Figure 7). The predominance of RZ errors may partly reflect the approximate axial symmetry of brain anatomy, although this interpretation has not been tested directly in the present experiment; the per-subject worst cases are all rotation-dominated. This diagnosis directly motivates the hybrid pipeline, whose iterative refinement targets precisely this residual rotational error.

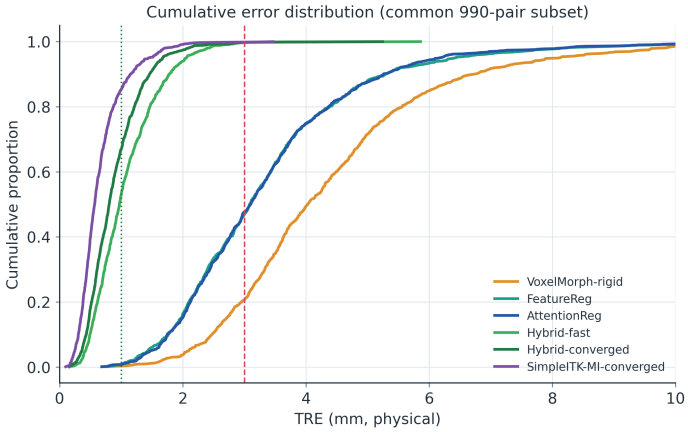


Figure 10. Cumulative distribution of the registration error in physical millimetres over the common 990-pair evaluation subset, so that every method is compared on identical cases. The guarded hybrid variants and the converged iterative reference lie clearly to the left of the standalone networks. The dotted and dashed lines mark the 1 mm and 3 mm references.

Because the 3 mm value is a convenient reference rather than a physically derived criterion, Figure 11 reports the success rate across a range of thresholds instead of a single one. The ordering of the three models is preserved at every threshold, and the margin

between the supervised models and VoxelMorph-rigid is widest in the sub-millimetre regime, where the choice of threshold matters most.

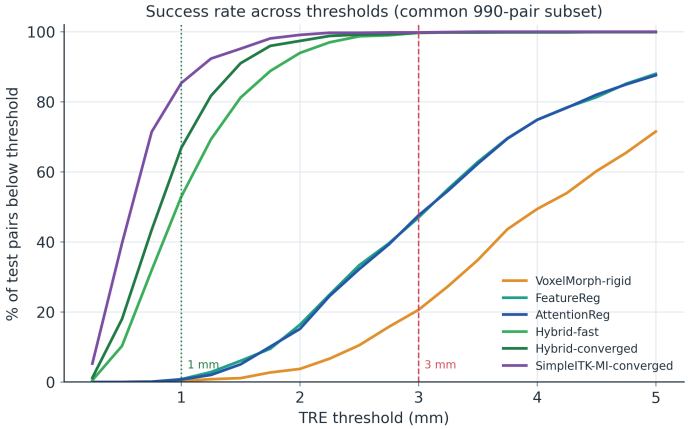


Figure 11. Proportion of test pairs registered below a given TRE threshold, for thresholds from 0.5 to 5 mm. Reporting the full curve rather than a single cut-off shows that the ranking of the models does not depend on the particular threshold chosen.

The accuracy also depends on how large the applied displacement is. Figure 12 stratifies the median TRE by quartile of applied transformation magnitude. All three models degrade as the displacement grows, but the two supervised models degrade more gently than VoxelMorph-rigid, and the gap between them widens in the upper quartile, which is the regime the failure-case analysis examines in detail (Section 3.6).

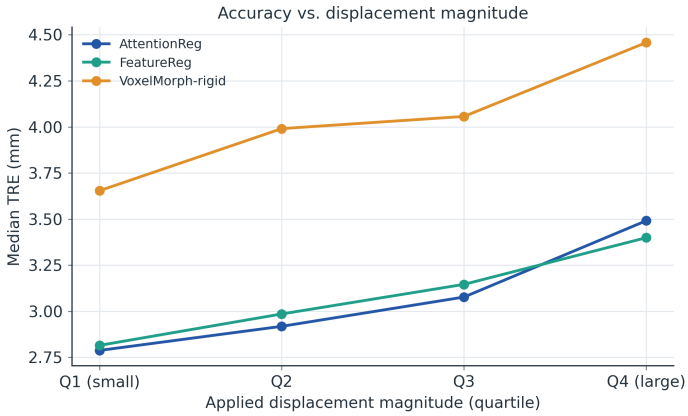


Figure 12. Median TRE stratified by quartile of applied transformation magnitude. Accuracy degrades with displacement for every model, most steeply for the compact baseline.

Table 4. Target Registration Error at the subject level on the common evaluation subset ($n = 99$; each subject contributes the mean over its 10 transformations). Values are physical millimetres; 95% confidence intervals are from 5000 bootstrap resamples over subjects.

Method	Median [95% CI]	Mean ± SD	% >3 mm
SimpleITK-MI-converged (reference)	0.61 [0.57, 0.67]	0.66 ± 0.24	0.0
VoxelMorph-rigid	4.21 [3.96, 4.49]	4.42 ± 1.26	99.0
FeatureReg (ablation)	3.18 [3.05, 3.32]	3.39 ± 1.11	67.7
AttentionReg (initialiser)	3.18 [3.08, 3.35]	3.38 ± 1.02	62.6
Hybrid-fast	**1.04 [0.95, 1.10]**	**1.07 ± 0.33**	**0.0**
Hybrid-converged (proposed)	**0.86 [0.79, 0.97]**	**0.89 ± 0.31**	**0.0**

Bold indicates the proposed guarded hybrid variants.

The ablation is decisive: AttentionReg and FeatureReg differ by −0.01 mm on the mean subject-level paired difference, with a confidence interval spanning zero and equivalence established at a 0.30 mm margin, whereas both supervised models outperform VoxelMorph-rigid by a large, highly significant margin. The principal gain therefore arises from end-to-end supervised regression rather than from the attention block per se. The specific contribution of attention is examined next.

3.4. The Three Learning Models at a Glance

Before the detailed error analysis, Figure 13 summarises the three learning-based models across five complementary axes and relates their capacity to their accuracy. AttentionReg and FeatureReg, which share the same 1.25M-parameter dual-path backbone, dominate on median accuracy, tail (P95) control, rotational accuracy, and robustness, while the far smaller VoxelMorph-rigid (0.16M parameters) is marginally faster. The proposed and ablation models are equivalent on the median but differ slightly on the rotational and tail axes, previewing the specific role of the attention block quantified below.

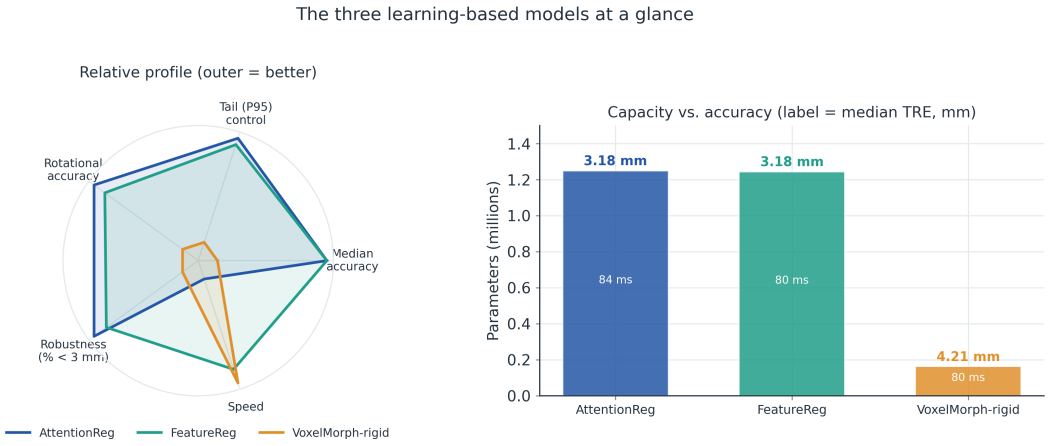


Figure 13. Simplified comparison of the three learning-based models. **Left**: relative performance profile (each axis normalised so that the outer edge is better) over median accuracy, tail control, rotational accuracy, robustness (proportion within 3 mm), and speed. **Right**: model capacity (parameters) against accuracy, annotated with median TRE and per-volume latency. The two dual-path models achieve markedly better accuracy at a modest capacity and latency cost relative to the compact VoxelMorph-rigid baseline.

3.5. Multimodal Specificity of the Attention Mechanism

Although the attention block does not change the median TRE, the per-parameter decomposition shows differences that are worth reporting, with the caveat that they are observations rather than statistically established improvements. Relative to the ablation, AttentionReg shows a lower mean absolute error on the rotational axes RX (0.680° vs. 0.728°) and RZ (1.051° vs. 1.092°), and a marginally higher error on the translations. Subject-level paired tests with Holm correction do not establish these differences: RX gives a corrected $p = 0.052$ with a confidence interval excluding zero, while RY ($p = 0.51$) and RZ ($p = 0.42$) have intervals spanning zero. We therefore describe the attention block as associated with a modest reduction in rotational and tail error, and do not claim a statistically demonstrated improvement.

3.6. Failure-Case Analysis

To characterise the residual error concretely, we examine the single worst subject in the test set. Subject 411 attains a mean TRE of 10.01 mm, roughly three times the subject-

level cohort median of 3.18 mm, and Figure 14 decomposes its hardest sample. Three points emerge.

First, the applied transformation lies at the boundary of the operating envelope established in Section 3.1: its largest rotation is 14.6° of a 15° bound (97%) and its largest translation is 9.2 of 10 resampled voxels (92%). The worst case is therefore not an anomaly but the expected consequence of sampling the extreme corner of the training distribution, which is exactly what the frame limit study was designed to delimit.

Second, the error is rotational. All three models recover the translations to within about 1.8 resampled voxels (4.5 mm), but diverge on rotation, and the axial rotation RZ is the principal offender: the applied value is 11.6°, and the models predict 7.3° (AttentionReg), 9.6° (FeatureReg) and 4.4° (VoxelMorph-rigid). This is the rotational ambiguity discussed above, expressed at its most severe. The pattern generalises: across the eight hardest subjects, the dominant error axis is rotational in every instance (RZ in six and RX in two), with no subject dominated by a translational axis.

Third, the ranking in this case is consistent with the role attributed to the attention block. AttentionReg attains the lowest rotational error of the three models (2.28° mean absolute, against 4.09° for the ablation and 3.29° for the baseline), while being marginally the weakest on translation (1.81 resampled voxels against 1.46 and 1.60). On the hardest case, in other words, attention buys rotational accuracy at a slight translational cost, the same trade-off visible in the aggregate per-parameter analysis.

This case also delimits what the standalone network can achieve, and motivates the hybrid pipeline directly: an initialisation carrying a 4° residual rotation remains well inside the convergence basin of a Mutual-Information optimiser, which is why the refinement stage recovers these cases (Section 3.8).

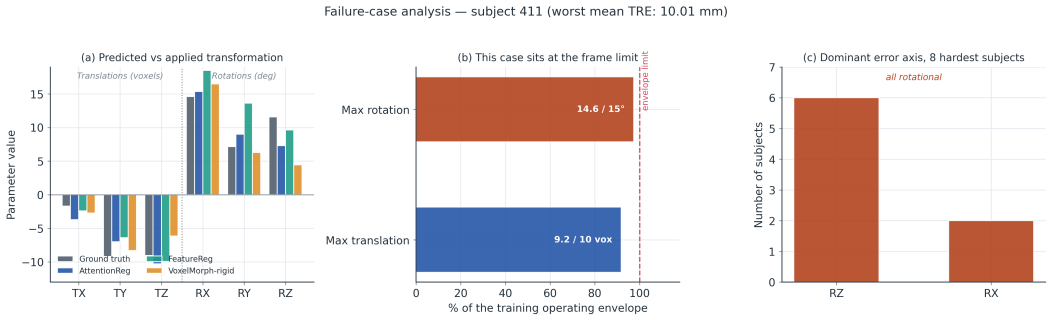


Figure 14. Failure-case analysis for subject 411, the worst subject in the test set. (**a**) Applied and predicted transformation parameters for the hardest sample; the models track the translations but diverge on the rotations, most visibly on RZ. (**b**) The applied transformation occupies 92% of the translational and 97% of the rotational training envelope, placing this case at the frame limit. (**c**) Dominant error axis across the eight hardest subjects: the failure mode is rotational in every case.

3.7. Comparison with Classical and Learning-Based Baselines

Table 5 reports every method on the common 990-pair subset using a single error measure in physical millimetres, and Figure 15 visualises the speed–accuracy trade-off. Three findings stand out.

First, only one classical configuration is accurate: the fully converged iterative reference reaches 0.57 mm, but requires 25.3 s per volume. Constraining the optimiser's budget destroys that accuracy entirely (SimpleITK-MI-fast: 11.21 mm, every case above 3 mm), and ANTs at default settings reaches 7.22 mm with 90.6% of cases above 3 mm. The local-minima problem of intensity-based multimodal registration is therefore severe when the optimiser is started from the identity transform and given a limited budget.

Second, the learning-based models are fast but not accurate enough to be used alone at this resolution. AttentionReg attains 3.10 mm and FeatureReg 3.08 mm, with roughly half of all pairs above the 3 mm threshold; VoxelMorph-rigid reaches 4.03 mm. This is a demanding regime for a single forward pass, and the values are reported here in physical units. At the 1 mm threshold, the standalone networks succeed in fewer than 1% of cases (Figure 11).

Third, and decisively, the guarded hybrid resolves the trade-off. It attains 0.81 mm (Hybrid-converged) and 0.97 mm (Hybrid-fast) with 0.3% of cases above 3 mm, at 0.79 and 0.92 s per volume, respectively. Relative to the converged reference, the difference is +0.23 mm (95% CI 0.19–0.28, Table 6). This difference is statistically significant, and we do not describe the two methods as indistinguishable; rather, two one-sided tests establish equivalence within a 0.50 mm margin, and the smallest margin at which equivalence holds is 0.28 mm. The hybrid therefore delivers accuracy statistically bounded within a third of a millimetre of the converged reference, in 1/32 of the time.

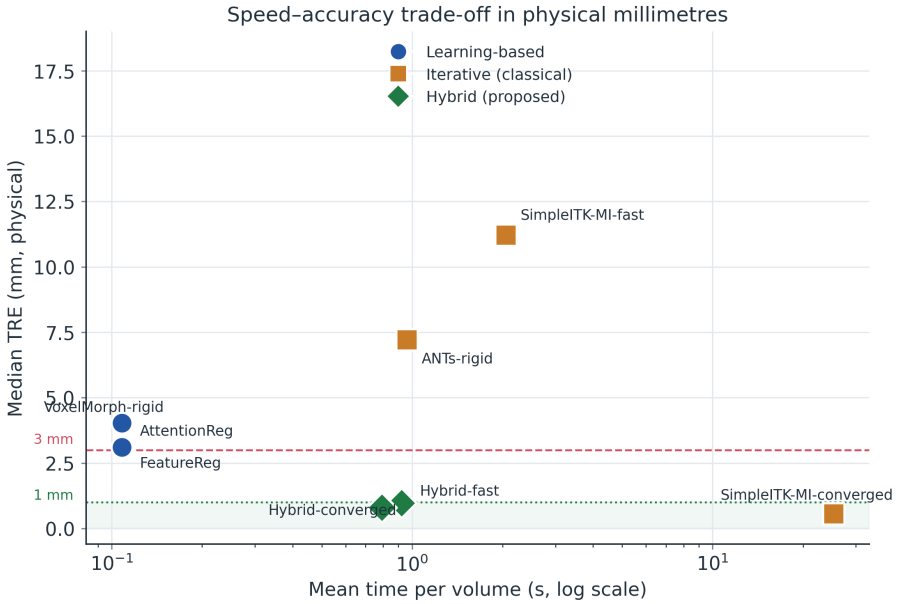


Figure 15. Speed–accuracy trade-off across all eight methods, in physical millimetres (mean time per volume on a logarithmic axis). Only three configurations fall below 1 mm: the two guarded hybrid variants and the fully converged iterative reference, which is more than thirty times slower. The dotted and dashed lines mark the 1 mm and 3 mm references.

Table 5. Complete comparison on the common evaluation subset (990 pairs from all 99 held-out subjects, 10 transformations each). Accuracy is the landmark-based Target Registration Error, identical in definition for every method. Values are given in physical millimetres and, in parentheses, in resampled-voxel units. Time is the mean per-volume registration time on identical hardware, recorded single-threaded for the iterative methods.

Method	Category	Median (mm)	P95 (mm)	Time (s)	% >3 mm
SimpleITK-MI-converged	iterative (converged)	0.57 (0.171)	1.45 (0.461)	25.28	0.2
SimpleITK-MI-fast	iterative (budgeted)	11.21 (2.538)	14.89 (4.195)	2.05	100.0
ANTs-rigid	iterative (default)	7.22 (2.270)	17.90 (5.014)	0.96	90.6
VoxelMorph-rigid	learning	4.03 (1.371)	8.11 (2.850)	0.108	79.4
FeatureReg (ablation)	learning	3.08 (1.033)	6.42 (2.289)	0.108	53.0
AttentionReg (initialiser)	learning	3.10 (1.024)	6.17 (2.121)	0.108	52.4
Hybrid-fast	**hybrid**	**0.97 (0.292)**	**2.06 (0.657)**	**0.923**	**0.3**
Hybrid-converged (proposed)	**hybrid**	**0.81 (0.251)**	**1.67 (0.546)**	**0.792**	**0.3**

Bold indicates the proposed guarded hybrid variants.

Table 6. Paired comparisons at the subject level ($n = 99$; each subject contributes the mean over its 10 transformations in the common subset). Wilcoxon signed-rank test, Cliff's δ, bootstrap 95% confidence interval of the mean paired difference (10,000 resamples), and two one-sided tests (TOST) for equivalence at the stated margin.

Comparison	Δ (mm) [95% CI]	p	δ	Margin	Equivalent
Hybrid-converged vs. converged reference	+0.23 [+0.19, +0.28]	1.1×10^{-15}	+0.50	0.50 mm	yes
Hybrid-fast vs. converged reference	+0.41 [+0.36, +0.46]	1.1×10^{-17}	+0.73	0.60 mm	yes
Hybrid-converged vs. AttentionReg	−2.49 [−2.67, −2.32]	5.7×10^{-18}	−1.00	—	no
AttentionReg vs. FeatureReg	−0.01 [−0.11, +0.09]	0.97	+0.01	0.30 mm	yes
AttentionReg vs. VoxelMorph-rigid	−1.05 [−1.22, −0.88]	1.0×10^{-16}	−0.66	—	no

3.8. Performance of the Guarded Hybrid Pipeline

Initialising the iterative optimiser with the AttentionReg prediction places it inside the convergence basin of the global optimum, so that only a local refinement remains. The effect is large: the median error falls from 3.10 mm for the network alone to 0.81 mm after refinement, corresponding to a mean subject-level paired reduction of 2.49 mm (95% CI 2.32–2.67, $\delta = -1.00$), which is a near-complete separation of the two distributions. Read across thresholds (Figure 11), the standalone network places 0.7% of pairs below 1 mm and 47.6% below 3 mm, whereas the hybrid places 66.9% below 1 mm and 99.7% below 3 mm.

The guard retains the refined transformation in 99% of cases. Because the guard operates on Mutual Information rather than on geometry, we measured how often an accepted refinement nonetheless increased the geometric error. It did so in 1.01% of cases for Hybrid-converged and 1.68% for Hybrid-fast, against a median improvement of 2.24 and 2.04 mm, respectively; the largest single degradation was 1.34 mm for Hybrid-converged and 7.73 mm for Hybrid-fast. The guard is therefore an effective but not absolute safeguard, and the residual risk it leaves is the natural place for the inference-time confidence estimate discussed in Section 4. Figure 16 illustrates the benefit on the six hardest cases for the ablation model. Refinement reduces the error substantially in every one, from a range of 12.6–17.2 mm to 2.1–11.7 mm, but two of the six remain above the 3 mm reference: these are the cases where the initialisation is furthest from the optimum and where a confidence flag would be most useful.

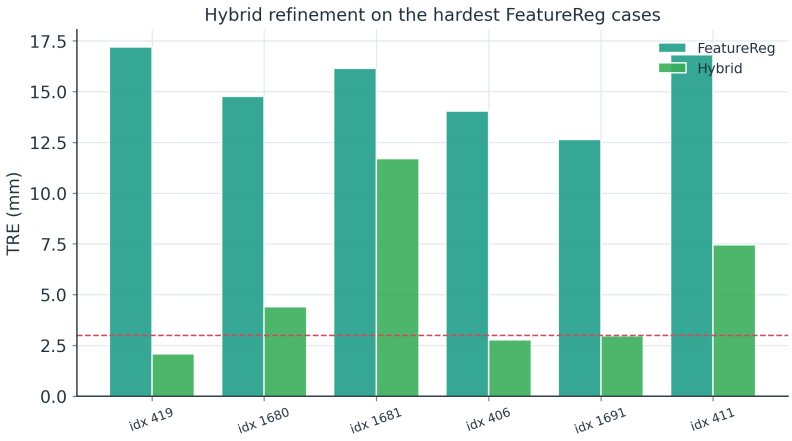


Figure 16. Warping index on the six hardest FeatureReg cases, before (FeatureReg) and after hybrid refinement. The red dotted line marks the 3 mm reference. Hybrid refinement reduces the error in every case and brings four of the six below the 3 mm reference, confirming that the iterative refinement targets the residual rotational error identified in Section 3.5.

Notably, Hybrid-converged is faster than Hybrid-fast (0.79 s vs. 0.92 s), the opposite of their standalone ordering. Because the initialisation is already near-optimal, the strict

convergence criterion of the converged optimiser triggers early termination almost immediately, whereas the fast configuration expends its fixed iteration budget. Figure 17 summarises the effect of initialisation on the refinement stage.

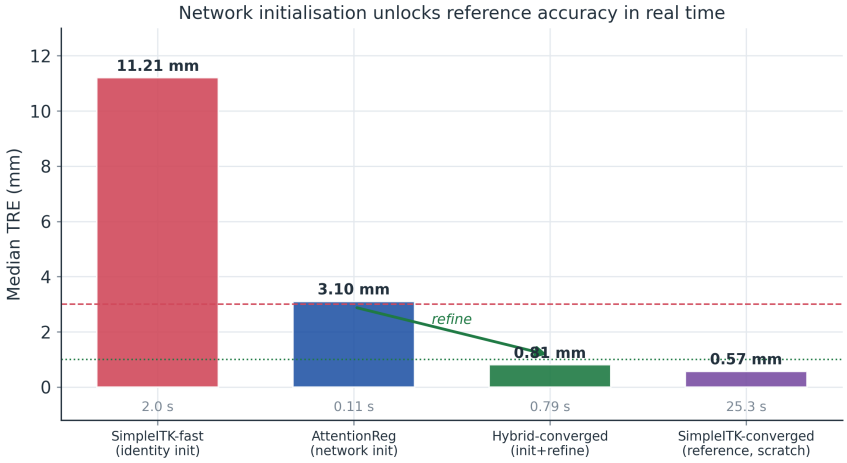


Figure 17. Effect of network initialisation on iterative refinement. Median error (bars) and mean time (annotations) for the fast iterative optimiser from an identity start, the network alone, the hybrid, and the fully converged iterative reference from scratch. The red dotted line marks the 3 mm reference. The network initialiser lets the hybrid reach near-reference accuracy in under one second.

4. Discussion

This study set out to determine whether a learned rigid registration predictor could be coupled with classical iterative optimisation to overcome the speed–accuracy trade-off that constrains multimodal MRI registration. The results support an affirmative and, in one respect, surprising answer. The following paragraphs interpret the principal findings and delimit the boundaries of the approach.

4.1. Breaking the Speed–Accuracy Trade off

The defining result is that the guarded hybrid attains a median error of 0.81 mm against 0.57 mm for the fully converged iterative reference, while reducing the per-volume time from 25.3 s to 0.79 s, a 32-fold acceleration, with 0.3% of cases exceeding the 3 mm reference. The difference of 0.23 mm is statistically significant but equivalent within a 0.50 mm margin. Neither family achieves this alone: the standalone learning models plateau near 3 mm because a single forward pass cannot perform the fine local search that sub-millimetre accuracy requires, whereas the budgeted iterative optimiser, started from an identity transform, is trapped in local minima created by the FLAIR–T1 contrast inversion (median 11.21 mm, every case above 3 mm). The learned predictor supplies precisely what the optimiser lacks: a starting point already inside the global convergence basin, so that only a short local refinement remains. This is a general and reusable pattern: a fast learned initialiser converts an accurate-but-slow iterative method into an accurate-and-fast one.

4.2. The Converged Faster than Fast Effect

A counter-intuitive observation reinforces this interpretation. The hybrid built on the *converged* optimiser is faster than the one built on the *fast* optimiser (0.79 s vs. 0.92 s), the opposite of their standalone ordering. Because the network initialisation is already near-optimal, the strict convergence criterion of the converged optimiser is satisfied after only a few micro-iterations and triggers early termination, whereas the fast configuration continues to expend its fixed iteration budget. When a strong initialiser is available, a strict early-stopping criterion is therefore both more accurate and faster than a loose fixed-budget one.

4.3. What the Attention Block Contributes

The controlled ablation shows that the cross-modal attention block does not change the median TRE relative to a matched dual-path CNN (negligible effect size, confidence interval spanning zero). The principal accuracy of the learned model thus derives from end-to-end supervised regression rather than from the attention mechanism in isolation. This is an honest and useful negative result, but it is not the whole picture: attention concentrates its benefit on the rotational parameters and on the tail of the error distribution, lowering the 95th-percentile TRE and the per-axis rotational error. In a pipeline whose residual error is rotation-dominated, and whose downstream iterative refinement is most fragile precisely where the initialisation is poorest, an initialiser that is more accurate and more stable on rotations is valuable even when its effect on the median is small.

4.4. Residual Error Is Rotational, Not Artefactual

Two independent lines of evidence localise the residual error of all learning-based models to rotation. During optimisation, the rotational component of the validation loss remains roughly six times the translational component for every architecture and throughout training (Figure 5); at test time, the per-parameter analysis confirms the same ordering and further localises the difficulty to the axial axis. This is consistent with a genuine rotational ambiguity arising from the near-symmetric appearance of the brain about the axial axis, and not with an acquisition-specific artefact. The hybrid pipeline resolves this residual by allowing the Mutual-Information optimiser to perform the fine rotational search that a single forward pass cannot, which is why the largest hybrid improvements occur on the cases where the standalone network error is greatest.

4.5. Processing Speed and Integration

Both the standalone network (84 ms) and the hybrid (under one second) are fast enough to be embedded as an inline step in a post-acquisition pipeline without measurably extending a reading session. Importantly, all timings are reported on identical hardware alongside the classical baselines, so the comparison reflects algorithmic rather than hardware differences. Where a single forward pass suffices, the network alone is appropriate; where sub-millimetre accuracy is required, the hybrid provides it at a cost still far below that of a converged iterative solve from scratch.

4.6. Relationship to Prior Learning-Based Registration

Most deep registration architectures (VoxelMorph [10], TransMorph [13], ViT-V-Net [14]) were designed for deformable registration and do not directly address rigid parameter estimation. Among rigid-specific methods, Song et al. [23] target MRI-to-ultrasound alignment, Sloan et al. [25] address monomodal rigid registration, and Chen et al. [26] focus on cardiac SPECT-CT. This work contributes a controlled, single-protocol comparison that includes a learning baseline adapted to the rigid setting and classical iterative baselines, and, distinctively, shows that the learned predictor is most valuable not as a standalone replacement but as an initialiser that unlocks real-time iterative refinement.

4.7. Limitations

Several limitations should be acknowledged. First, all training and test pairs are generated by synthetic rigid transformation of pre-aligned volumes rather than from acquisitions with concurrently recorded motion; the extent to which synthetic displacement captures real motion statistics (through-plane motion, cardiac pulsation, gradient-induced eddy currents) remains to be established through prospective study. Second, the volumes are downsampled to $96 \times 96 \times 32$ voxels to respect memory constraints, which discards fine

detail; higher-resolution training would likely reduce residual error. Third, the evaluation is confined to the publicly available BraTS cohort under a synthetic-transform protocol; validation on multi-site clinical acquisitions with real motion is the natural next step and would benefit from an inference-time confidence estimate to flag the rare high-uncertainty cases for review. Finally, the hybrid inherits the assumptions of its iterative component, and one of these deserves to be made explicit. The refinement stage restricts its Mutual-Information metric to a brain mask, which on BraTS is trivially obtained because the volumes are distributed skull-stripped: the background is exactly zero and a simple intensity threshold recovers the brain. On raw clinical acquisitions, this is no longer the case. Skull, scalp and background noise are present, so the same threshold would retain non-cerebral tissue, and the mask would no longer isolate the anatomy the metric is meant to compare. Deploying the hybrid outside a skull-stripped setting therefore requires an explicit brain-extraction step, or a mask definition robust to its absence. More generally, the iterative component assumes a degree of intensity regularity that clinical data do not guarantee; bias-field inhomogeneity in particular distorts the joint histogram on which Mutual Information is computed, so intensity correction is likely to become a prerequisite rather than an option. We expect the hybrid to degrade more gracefully than the network alone, since an initialisation that loses accuracy remains within the convergence basin the refinement needs, but robustness to preprocessing variation across sites should be characterised before deployment.

5. Conclusions

This work shows that a fast learned predictor and a classical iterative optimiser, combined in a guarded hybrid pipeline, resolve the speed–accuracy trade-off that constrains multimodal rigid MRI registration. On a common subset of 990 held-out pairs from 99 subjects, evaluated with a single landmark-based error measure expressed in physical millimetres, the proposed pipeline attains a median Target Registration Error of 0.81 mm (P95 1.67 mm; 0.3% of cases above 3 mm) at 0.79 s per volume. The fully converged iterative reference attains 0.57 mm but requires 25.3 s. The paired difference is 0.23 mm (95% CI 0.19–0.28): statistically significant, and equivalent within a 0.50 mm margin, for a 32-fold reduction in computation time.

The decisive observation concerns the role of the learned component. Used alone, the network reaches 3.10 mm and places fewer than 1% of pairs below 1 mm, which is not sufficient for the intended use. Its value lies in initialisation: supplying the optimiser with a starting point inside the global convergence basin reduces the error by a mean subject-level paired reduction of 2.49 mm (95% CI 2.32–2.67) and converts a method that is accurate only after 25 s into one that is accurate in under a second. A controlled ablation shows that this capability does not depend on the cross-modal attention block, which leaves median accuracy unchanged relative to a matched network without it; the attention block is associated with a modest reduction in rotational and tail error that the present sample does not establish statistically.

Two practical guidelines follow. Once a strong initialisation is available, a strict early-stopping optimiser becomes both more accurate and faster than a fixed-budget one. And a guard defined on an intensity metric protects against most, but not all, geometric degradation: it accepted a refinement that increased the error in 1% of cases, which delimits precisely where an inference-time confidence estimate would be useful.

Taken together, these findings establish a proof of concept for real-time rigid multimodal MRI registration at an accuracy previously reachable only at a computational cost incompatible with interactive use. Future work will pursue higher-resolution and isotropic resampling, which the present anisotropic grid limits; add an inference-time confidence

estimate to flag the residual high-uncertainty cases; and validate on multi-site clinical acquisitions with concurrently recorded ground-truth motion.

Author Contributions: Conceptualisation, Z.S.; methodology, Z.S.; software, Z.S.; formal analysis, Z.S., F.-E.B.-B. and M.M.; investigation, Z.S.; data curation, Z.S.; writing original draft preparation, Z.S.; writing review and editing, F.-E.B.-B. and M.M.; visualisation, Z.S.; supervision, F.-E.B.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study uses only the publicly available, fully anonymised BraTS 2020 dataset. No private patient data were used and no new human or animal data were collected, so institutional review board approval was not required.

Informed Consent Statement: Not applicable.

Data Availability Statement: The BraTS 2020 dataset is publicly available at https://www.med.upenn.edu/cbica/brats2020/data.html (accessed on 16 September 2026). All derived results reported in this article, including per-parameter errors, per-sample registration errors, baseline and hybrid outputs, and the summary statistics underlying every table and figure, are contained within the article. The implementation is research code that has not been prepared for public release: it is incompletely documented and retains development comments, so releasing it in its present state would not serve reproducibility. The code is available from the corresponding author on reasonable request.

Acknowledgments: The authors thank the BraTS 2020 challenge organisers for open access to the dataset. During the preparation of this manuscript, the authors used Claude (Anthropic) to improve the language and phrasing of some paragraphs. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Yang, X.; Wu, N.; Cheng, G.; Zhou, Z.; Yu, D.S.; Beitler, J.J.; Curran, W.J.; Liu, T. Automated segmentation of the parotid gland based on atlas registration and machine learning: A longitudinal MRI study in head-and-neck radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **2014**, *90*, 1225–1233. [CrossRef] [PubMed]
2. Chandrashekara, R.; Rao, A.; Sanchez-Ortiz, G.I.; Mohiaddin, R.H.; Rueckert, D. Construction of a statistical model for cardiac motion analysis using nonrigid image registration. In *Information Processing in Medical Imaging*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 599–610.
3. Yang, X.; Akbari, H.; Halig, L.; Fei, B. 3D non-rigid registration using surface and local salient features for transrectal ultrasound image-guided prostate biopsy. *Proc. SPIE* **2011**, *7964*, 79642V. [CrossRef] [PubMed]
4. Fu, Y.; Lei, Y.; Wang, T.; Curran, W.J.; Liu, T.; Yang, X. Deep learning in medical image registration: A review. *Phys. Med. Biol.* **2020**, *65*, 20TR01. [CrossRef] [PubMed]
5. Brown, L.G. A survey of image registration techniques. *ACM Comput. Surv.* **1992**, *24*, 325–376. [CrossRef]
6. Xiao-chun, Z.; Xin-bo, Z.; Yan, F. An efficient medical image registration algorithm based on gradient descent. In Proceedings of the IEEE/ICME International Conference on Complex Medical Engineering, Beijing, China, 23–27 May 2007; pp. 636–639.
7. Oliveira, F.P.M.; Tavares, J.M.R.S. Novel framework for registration of pedobarographic image data. *Med. Biol. Eng. Comput.* **2011**, *49*, 313–323. [CrossRef] [PubMed]
8. Zagoruyko, S.; Komodakis, N. Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4353–4361.
9. Salehi, S.S.M.; Khan, S.; Erdogmus, D.; Gholipour, A. Real-time deep registration with geodesic loss. *arXiv* **2018**, arXiv:1803.05982.
10. de Vos, B.D.; Berendsen, F.F.; Viergever, M.A.; Sokooti, H.; Staring, M.; Išgum, I. A deep learning framework for unsupervised affine and deformable image registration. *Med. Image Anal.* **2019**, *52*, 128–143. [CrossRef] [PubMed]
11. Chee, E.; Wu, Z. AIRNet: Self-supervised affine registration for 3D medical images using neural networks. *arXiv* **2018**, arXiv:1810.02583.
12. Liao, R.; Miao, S.; de Tournemire, P.; Grbic, S.; Kamen, A.; Mansi, T.; Comaniciu, D. An artificial agent for robust image registration. *arXiv* **2016**, arXiv:1611.10336.

13. Chen, J.; Frey, E.C.; He, Y.; Segars, W.P.; Li, Y.; Du, Y. TransMorph: Transformer for unsupervised medical image registration. *Med. Image Anal.* **2022**, *82*, 102615. [CrossRef] [PubMed]
14. Chen, J.; He, Y.; Frey, E.C.; Li, Y.; Du, Y. ViT-V-Net: Vision transformer for unsupervised volumetric medical image registration. *arXiv* **2021**, arXiv:2104.06468.
15. Mustafa, W.; Ali, S.; Elgendy, N.; Salama, S.; El Sorogy, L.; Mohsen, M. Role of contrast-enhanced FLAIR MRI in diagnosis of intracranial lesions. *Egypt. J. Neurol. Psychiatry Neurosurg.* **2021**, *57*, 108. [CrossRef]
16. Roozpeykar, S.; Azizian, M.; Zamani, Z.; Farzan, M.R.; Veshnavei, H.A.; Tavoosi, N.; Toghyani, A.; Sadeghian, A.; Afzali, M. Contrast-enhanced weighted-T1 and FLAIR sequences in MRI of meningeal lesions. *Am. J. Nucl. Med. Mol. Imaging* **2022**, *12*, 63–70. [PubMed]
17. Sati, P.; George, I.C.; Shea, C.D.; Gaitán, M.I.; Reich, D.S. FLAIR*: A combined MR contrast technique for visualising white matter lesions and parenchymal veins. *Radiology* **2012**, *265*, 926–932. [CrossRef] [PubMed]
18. Avants, B.B.; Tustison, N.J.; Song, G.; Cook, P.A.; Klein, A.; Gee, J.C. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage* **2011**, *54*, 2033–2044. [CrossRef] [PubMed]
19. Menze, B.H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans. Med. Imaging* **2015**, *34*, 1993–2024. [CrossRef] [PubMed]
20. Bakas, S.; Akbari, H.; Sotiras, A.; Bilello, M.; Rozycki, M.; Kirby, J.S.; Freymann, J.B.; Farahani, K.; Davatzikos, C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data* **2017**, *4*, 170117. [CrossRef] [PubMed]
21. Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R.T.; Berger, C.; Ha, S.M.; Rozycki, M.; et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv* **2018**, arXiv:1811.02629.
22. Lowekamp, B.C.; Chen, D.T.; Ibáñez, L.; Blezek, D. The Design of SimpleITK. *Front. Neuroinform.* **2013**, *7*, 45. [CrossRef] [PubMed]
23. Song, X.; Chao, H.; Xu, X.; Guo, H.; Xu, S.; Turkbey, B.; Wood, B.J.; Sanford, T.; Wang, G.; Yan, P. Cross-modal attention for multi-modal image registration. *Med. Image Anal.* **2022**, *82*, 102612. [CrossRef] [PubMed]
24. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
25. Sloan, J.M.; Goatman, K.A.; Siebert, J.P. Learning rigid image registration Utilising convolutional neural networks for medical image registration. In Proceedings of the Bioimaging 2018, Funchal, Portugal, 19–21 January 2018.
26. Chen, X.; Zhou, B.; Xie, H.; Guo, X.; Zhang, J.; Sinusas, A.J.; Onofrey, J.A.; Liu, C. Dual-branch squeeze-fusion-excitation module for cross-modality registration of cardiac SPECT and CT. In *Proceedings of the Medical Image Computing and Computer Assisted Intervention (MICCAI 2022), Singapore, 18–22 September 2022*; Springer: Cham, Switzerland, 2022; pp. 46–55.