SPEAR: Error-Guided Sparse Refinement for Scalable Deformable Medical Image Registration

Hao Bai and Yi Hong*

School of Computer Science, Shanghai Jiao Tong University, Shanghai, China
yi.hong@sjtu.edu.cn

Abstract. Accurate and efficient deformable medical image registration remains challenging, particularly for high-resolution 3D volumes where dense multi-scale architectures incur substantial computational and memory overhead. We propose **SPEAR**, a scalable error-guided sparse refinement framework for 3D deformable registration. It leverages anatomy-aware discrepancy signals to guide targeted sparse refinement, enabling progressive deformation updates without redundant computation. Experiments show that SPEAR improves deformation accuracy and regularity over conventional unsupervised models, while achieving comparable performance to recent multi-scale approaches with substantially lower computational and memory cost. Importantly, the proposed design scales effectively to high-resolution 3D volumes under constrained GPU memory. These results demonstrate that structured sparse refinement offers a scalable and balanced solution for deformable medical image registration. Code: https://github.com/ZedKing12138/SPEAR-registration.

Keywords: Deformable Image Registration · Sparse Refinement · Scalable 3D Modeling

1 Introduction

Deformable medical image registration aims to establish dense voxel-wise anatomical correspondences between volumetric scans and is fundamental to longitudinal analysis [11, 3], atlas construction [1, 8], and image-guided intervention [9]. Learning-based registration frameworks have significantly improved efficiency over traditional optimization-based methods. Unsupervised approaches such as VoxelMorph [2] directly optimize image similarity and deformation smoothness, enabling fast inference without ground-truth deformation fields. Transformer-based extensions like TransMorph [4] enhance global correspondence modeling through attention mechanisms. However, they rely on dense convolution or attention that uniformly operates over the full volume, leading to inefficient modeling of spatially varying local deformations, especially at high resolution.

To improve alignment precision, recent works adopt multi-scale coarse-to-fine strategies. Representative approaches such as CorrMLP [14] and SACB-Net [5] progressively estimate deformation fields across hierarchical resolutions.

* Corresponding author.

Although multi-scale coarse-to-fine strategies improve alignment accuracy, they introduce redundant cross-scale computation and high memory overhead. These limitations are amplified in high-resolution 3D volumes, where costs scale rapidly with spatial resolution. To satisfy GPU constraints, such models often rely on aggressive downsampling, potentially compromising fine anatomical alignment. This calls for a refinement mechanism that selectively focuses computation on anatomically misaligned regions without redundant dense processing.

Inspired by token stacking strategies in large vision models [13], we propose **SPEAR**, a scalable error-guided sparse refinement framework for 3D deformable registration. Instead of dense multi-scale pyramids, it performs targeted progressive refinement via structured sparse feature injection. Anatomical misalignment is estimated using segmentation-derived discrepancy signals, which guide region-wise Top-r routing for selective local refinement under a non-redundant constraint. SPEAR improves deformation accuracy and regularity over conventional unsupervised models, while matching recent coarse-to-fine approaches with significantly lower computational overhead. Its structured sparse design enables scalable high-resolution 3D registration under constrained GPU memory.

The main contributions are summarized as follows.

- We propose SPEAR, a scalable error-guided sparse refinement framework that replaces dense multi-scale hierarchies with targeted region-wise updates.
- We introduce a strictly non-redundant routing mechanism that enforces complementary local exploration across refinement stages, improving deformation stability while avoiding redundant computation.
- We demonstrate that SPEAR achieves a favorable balance between registration accuracy, deformation regularity, and computational efficiency, scaling effectively to high-resolution 3D volumes under constrained GPU memory.

2 Method

Problem Formulation. Let $\Omega \subset \mathbb{R}^3$ denote the 3D spatial domain. Given a moving image $I_m : \Omega \to \mathbb{R}$ and a fixed image $I_f : \Omega \to \mathbb{R}$, deformable registration aims to estimate a dense deformation $\phi : \Omega \to \Omega$ such that the warped image $I_m \circ \phi = I_m(\phi(x))$ is spatially aligned with I_f. Instead of directly regressing ϕ in one step, we adopt a progressive refinement that decomposes large and complex deformations into a sequence of residual updates.

Framework Overview. Figure 1 illustrates the overall architecture of SPEAR. Moving and fixed images are first encoded by a shared 3D encoder into hierarchical token representations. Building on this, SPEAR performs a progressive refinement over N stages. At each stage, a region-wise Top-r sparse routing mechanism identifies informative token correspondences under a hard non-redundant constraint. The selected tokens are then injected into a cross-attention transformer to predict a residual displacement field, which is composed with previous updates to iteratively refine the alignment. To further guid the refinement, an error map computed from the current warped image steers token routing in subsequent stages, forming an error-driven feedback loop. Notably, sparsity is imposed only

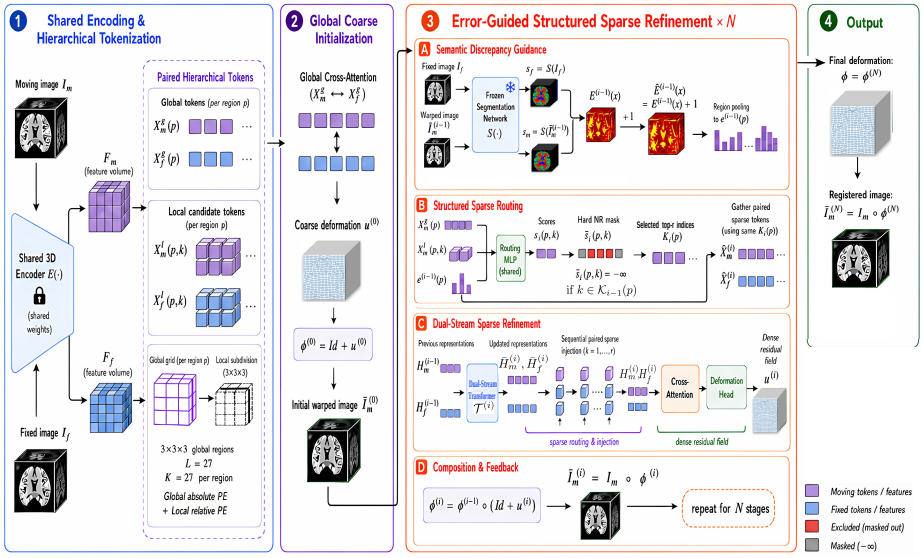

Fig. 1. Overview of our method. SPEAR first estimates a coarse deformation from paired global tokens, then progressively refines it using semantic-discrepancy-guided Top-r sparse routing and token injection. Each stage predicts a dense residual displacement, which is composed into the final deformation. Training uses only image similarity and smoothness losses, without segmentation or label-overlap supervision.

on token routing and injection, while the residual deformation field is predicted densely to ensure spatial continuity and smoothness.

Hierarchical Tokenization. A shared encoder $\mathcal{E}$ extracts latent feature maps $F_m = \mathcal{E}(I_m)$ and $F_f = \mathcal{E}(I_f)$, where $F_m, F_f \in \mathbb{R}^{C\times H\times W\times D}$ denote feature tensors with C channels and spatial resolution $H \times W \times D$. Since global attention scales quadratically with the number of tokens [16], directly constructing dense voxel-wise tokens for 3D volumes is computationally expensive. We therefore organize the feature embeddings into a two-level hierarchy of global and local tokens to balance representation capacity and computational efficiency.

Global Tokens. The feature domain is evenly partitioned into L non-overlapping volumetric regions $\{\Omega_p\}_{p=1}^{L}$. In practice, we adopt a $3\times3\times3$ grid, yielding $L = 27$ global regions. This coarse partition defines structured search units for large-scale anatomical correspondence. For each region Ω_p, a pooled feature vector is computed as $X_m^g(p) = \mathrm{Pool}(F_m|_{\Omega_p})$, $X_f^g(p) = \mathrm{Pool}(F_f|_{\Omega_p})$, where $\mathrm{Pool}(\cdot)$ denotes average pooling. Global tokens summarize coarse anatomical context and serve as structured anchors for region-wise sparse routing, enabling large-scale correspondence modeling before fine-grained refinement.

Local Candidate Tokens. Within each global region Ω_p, we further subdivide it into K finer subregions $\{\Omega_{p,k}\}_{k=1}^{K}$, with $K = 27$ in our implementation. Local candidate tokens are defined as $X_m^\ell(p,k) = \mathrm{Pool}(F_m|_{\Omega_{p,k}}), X_f^\ell(p,k) =$

$\mathrm{Pool}(F_f|_{\Omega_{p,k}})$. All global and local candidate tokens are projected to a unified embedding dimension d via a shared linear layer before cross-attention. This hierarchical organization restricts attention to structured local neighborhoods, enabling fine-grained refinement without constructing explicit multi-scale pyramids. Compared with flat tokenization, it reduces redundant comparisons and provides an explicit coarse-to-fine matching prior.

To preserve spatial correspondence, we incorporate two positional schemes. *(1) Global Absolute Encoding.* Global tokens use 3D sinusoidal encoding based on the normalized centroid of each region Ω_p, providing spatial awareness for region-level routing. *(2) Local Relative Encoding.* Local candidate tokens employ learnable relative offset embeddings to encode their positions within each global region. Together, these encodings maintain spatial coherence and anatomical consistency during sparse routing.

Global Coarse Initialization. Before sparse refinement, we first estimate a coarse global deformation using only global tokens. The cross-attention of moving and fixed global tokens produces an initial residual displacement $u^{(0)}$, yielding $\phi^{(0)} = \mathrm{Id} + u^{(0)}$ (Id is the identity map). Subsequent stages focus on structured local refinement guided by discrepancy signals.

Error-Guided Structured Sparse Refinement. Although hierarchical tokenization constrains correspondence within regions, refining all local candidates remains redundant. We therefore introduce region-wise sparse routing. At stage i, each local candidate k in region p is assigned a routing score: $s_i(p,k) = \mathrm{MLP}\big(X_m^g(p), X_m^\ell(p,k), e^{(i-1)}(p)\big)$, where $X_m^g(p)$ and $X_m^\ell(p,k)$ denote the global and local tokens, and $e^{(i-1)}(p)$ is the aggregated discrepancy signal. The MLP outputs a scalar importance score for refinement.

Semantic Discrepancy Guidance. Sparse routing requires a reliable indicator of anatomical misalignment. Intensity differences are often sensitive to appearance variation and noise. We therefore introduce a segmentation-guided discrepancy signal. A pretrained segmentation network $\mathcal{S}$, trained on the training split of each dataset, is used as a frozen anatomical feature extractor.

At stage $i-1$, we compute the warped image $\tilde{I}_m^{(i-1)} = I_m \circ \phi^{(i-1)}$. Soft segmentation predictions are obtained as $s_f = \mathcal{S}(I_f)$, $s_m = \mathcal{S}(\tilde{I}_m^{(i-1)})$, where $s_f, s_m \in \mathbb{R}^{C_s \times H \times W \times D}$ denote soft class-probability maps with C_s semantic classes, and $H \times W \times D$ matches the spatial resolution of the input volumes. *(1) Voxel-wise Discrepancy.* We define the semantic discrepancy at voxel $\mathbf{x}$ as $E^{(i-1)}(\mathbf{x}) = \frac{1}{C_s}\sum_{c=1}^{C_s}\left\|s_f^c(\mathbf{x}) - s_m^c(\mathbf{x})\right\|_2^2$. This formulation measures class-probability disagreement, capturing structural misalignment beyond raw intensity differences. To avoid vanishing routing signals in well-aligned regions and ensure stable aggregation across stages, we apply a constant positive shift $\hat{E}^{(i-1)}(\mathbf{x}) = E^{(i-1)}(\mathbf{x}) + 1$. *(2) Region-level Aggregation.* Region-wise discrepancy is obtained via pooling: $e^{(i-1)}(p) = \mathrm{Pool}_{\mathbf{x}\in\Omega_p}\hat{E}^{(i-1)}(\mathbf{x})$. The resulting region-level signal highlights anatomically inconsistent areas and serves as a guidance factor for structured sparse routing in the next stage.

Structured Sparse Routing. For each global region p, we select the top-r local tokens, i.e., $\mathcal{K}_i(p) = \text{Top-}r\big(s_i(p, 1 : K)\big)$. The total number of refined tokens per stage is $L \times r$ (i.e., $27r$ in our setting). This structured sparsity reduces local refinement from $L \times K$ candidates to $L \times r$, yielding a $\frac{K}{r}$-fold reduction in computational complexity. Unlike global Top-r selection, which may concentrate refinement on a few dominant regions, this region-wise strategy guarantees that each global region retains r refinement opportunities per stage. This prevents spatial collapse and maintains uniform modeling capacity across the entire volume. Although region-wise sparse routing reduces redundancy within a stage, repeatedly selecting the same subregions across consecutive stages may lead to oscillatory updates and inefficient refinement allocation. To promote progressive coverage and stabilize refinement dynamics, we introduce a hard non-redundant constraint between stages. Let $\mathcal{K}_{i-1}(p)$ denote the set of selected local indices in global region p at stage $i-1$. During stage i, these indices are masked: $\tilde{s}_i(p, k) = -\infty$ if $k \in \mathcal{K}_{i-1}(p)$; otherwise, it is $s_i(p, k)$. The Top-r selection is performed on $\tilde{s}_i(p, k)$.

This constraint enforces temporal diversity in refinement, encouraging the model to progressively explore different subregions instead of repeatedly correcting the same local areas. Importantly, the restriction is applied only between consecutive stages, allowing subregions to be reconsidered after sufficient global deformation updates. By preventing redundant local corrections, the proposed mechanism improves refinement efficiency and mitigates local update oscillation.

Dual-Stream Sparse Refinement. The selected sparse candidate tokens are injected into a symmetric dual-stream transformer backbone to progressively update deformation representations. The moving and fixed streams evolve jointly across stages while preserving a fixed number of global tokens. Let $X_m^g, X_f^g \in \mathbb{R}^{L\times C}$ denote the global tokens extracted from the moving and fixed images. We initialize $H_m^{(0)} = X_m^g$ and $H_f^{(0)} = X_f^g$. At stage i, the previous representations are first updated by a dual-stream transformer block: $\bar{H}_m^{(i)}, \bar{H}_f^{(i)} = \mathcal{T}^{(i)}(H_m^{(i-1)}, H_f^{(i-1)})$. Region-wise sparse routing then selects a Top-r candidate set per region, from which sparse tokens $\tilde{X}_m^{(i)}, \tilde{X}_f^{(i)} \in \mathbb{R}^{L\times r\times C}$ are gathered. Instead of concatenating all candidates, we adopt *progressive sparse injection*. For $k = 1, \ldots, r$, the streams are updated sequentially:

$$H_{m,k}^{(i)} = \mathcal{P}_{m,k}^{(i)}(H_{m,k-1}^{(i)}) + \tilde{X}_{m,k}^{(i)}, \quad H_{f,k}^{(i)} = \mathcal{P}_{f,k}^{(i)}(H_{f,k-1}^{(i)}) + \tilde{X}_{f,k}^{(i)}, \tag{1}$$

where $\tilde{X}_{(\cdot),k}^{(i)} \in \mathbb{R}^{L\times C}$ denotes the k-th selected token (one per region). This sequential injection preserves a constant token dimension and avoids quadratic growth in token interactions. After r sub-steps, we obtain $H_m^{(i)}$ and $H_f^{(i)}$. A stage-level cross-attention layer then models inter-stream correspondences: $Z^{(i)} = \text{CrossAttn}(H_m^{(i)}, H_f^{(i)})$, followed by a lightweight regression head that predicts the residual displacement $u^{(i)} = \mathcal{H}(Z^{(i)})$.

Composition and Feedback. SPEAR composes N residual displacement fields as $\phi^{(i)} = \phi^{(i-1)} \circ (\text{Id} + u^{(i)})$, $i = 1, \ldots, N$. After each update, the warped moving

image is compared with the fixed image to produce an error map, which is fed back to guide sparse routing in the next stage. This feedback loop lets subsequent stages focus on the remaining misalignment, yielding the final deformation $\phi = \phi^{(N)}$. This recursive composition supports coarse-to-fine deformation modeling and stable optimization under large anatomical variations.

Training Objective. We optimize the network using an image-based registration objective. The segmentation network $\mathcal{S}$ used for discrepancy estimation remains frozen, and no anatomical labels or overlap losses directly supervise the deformation field. The registration network is therefore optimized by image similarity and deformation smoothness, while semantic predictions are used only to allocate sparse refinement. Hence, the overall objective is defined as

$$\mathcal{L} = \sum_{i=1}^{N} \lambda_i \mathcal{L}_{\text{sim}}\big(I_f, I_m \circ \phi^{(i)}\big) + \mu \sum_{i=1}^{N} \mathcal{L}_{\text{smooth}}(u^{(i)}), \tag{2}$$

where λ_i balances progressive supervision across refinement stages. Similarity $\mathcal{L}_{\text{sim}}$ is measured using normalized cross-correlation (NCC) [2], and smoothness is enforced via spatial gradient regularization: $\mathcal{L}_{\text{smooth}}(u^{(i)}) = \|\nabla u^{(i)}\|_2^2$.

3 Experiments

Datasets and Experimental Settings. We evaluate SPEAR on three public T1-weighted brain MRI datasets: IXI (581 scans), OASIS (416 scans), and IBSR18 (18 high-resolution volumes), covering diverse anatomical variability and spatial resolutions [7, 12, 15]. Images are affinely pre-aligned and intensity normalized. To support the two-level $3 \times 3 \times 3$ partitioning, volumes are resampled to resolutions divisible by 9: 144^3 (IXI), $162 \times 198 \times 162$ (OASIS), and 225^3 (IBSR18). IBSR18 retains larger spatial dimensions for scalability evaluation.

All experiments follow an inter-subject pairwise registration setting with a 7:1:2 split for training, validation, and testing, ensuring no subject overlap. Within each subject-wise split, we construct all ordered non-self moving–fixed pairs. For the four IBSR18 test subjects, this results in $4 \times (4-1) = 12$ pairwise evaluations. Notably, training annotations are used solely to pretrain the frozen auxiliary segmenter, while test annotations are reserved exclusively for Dice evaluation; neither is involved in optimizing the registration objective. Deformation regularity is measured by folding rate (percentage of voxels with $\det(\nabla\phi) \leq 0$). Efficiency is assessed by inference time, peak GPU memory, and FLOPs under the stated evaluation protocol.

Implementation Details and Baselines. Experiments are conducted on an NVIDIA L40S GPU (48GB). We use $N = 3$ refinement stages with Top-$r = 6$ region-wise routing. A lightweight 3D U-Net [6] is pretrained on the training split of each dataset using cross-entropy loss and frozen during registration training. The shared encoder preserves spatial resolution while progressively increasing channel dimensions. The dual-stream transformer contains 4 layers with 8 attention heads. The deformation head $\mathcal{H}$ consists of two $3 \times 3 \times 3$ convolutions pre-

Table 1. Inter-subject registration results on three public datasets.

Dataset	Method	Dice ↑	Foldings (%) ↓	Time (s) ↓	Mem(GB)↓	FLOPs(G)↓
IXI	VoxelMorph	0.743±0.130	0.380±0.250	**0.2364**	**9.8**	**305.36**
	TransMorph	0.761±0.128	0.785±0.276	0.2538	20.4	396.27
	CorrMLP	0.791±0.133	0.158±0.096	0.9261	33.6	601.14
	SACB-Net	**0.795±0.114**	**0.089±0.032**	1.1563	36.9	673.22
	SPEAR (ours)	0.785±0.169	0.112±0.061	0.3492	25.7	423.58
OASIS	VoxelMorph	0.722±0.127	0.394±0.216	**0.2834**	**12.7**	**384.38**
	TransMorph	0.738±0.112	0.811±0.243	0.3218	26.5	451.22
	CorrMLP	0.769±0.156	0.151±0.080	1.1260	37.9	681.10
	SACB-Net	**0.778±0.166**	**0.095±0.041**	1.2613	44.3	760.23
	SPEAR (ours)	0.753±0.143	0.120±0.065	0.4792	30.2	535.18
IBSR18	VoxelMorph	0.659±0.023	0.452±0.224	0.6431	**23.6**	502.18
	TransMorph	0.696±0.038	0.921±0.452	0.6843	36.8	496.27
	CorrMLP	0.614±0.048	1.357±0.622	0.7562	30.2	**431.49**
	SACB-Net	0.619±0.043	1.716±0.735	0.9317	35.7	468.15
	SPEAR (ours)	**0.722±0.029**	**0.172±0.051**	**0.5129**	30.9	455.46

dicting residual displacement. Models are trained for 500 epochs using Adam [10] (initial learning rate 1×10^{-4}, halved every 150 epochs) with batch size 1.

We compare against four unsupervised registration baselines: VoxelMorph [2], TransMorph [4], CorrMLP [14], and SACB-Net [5]. On IXI and OASIS, all methods are trained at full resolution. On IBSR18, CorrMLP and SACB-Net exceed memory limits at full resolution; thus inputs are downsampled by 2× and their predicted flows are upsampled to 225^3 before warping labels and computing Dice. Reported runtime, memory, and FLOPs include resampling overhead. Therefore, IBSR18 is interpreted as a fixed-memory scalability evaluation rather than a strictly same-input-resolution accuracy comparison. All baselines are trained using official implementations with identical data splits.

Experimental Results. Table 1 reports quantitative results on our datasets. Compared with conventional unsupervised models, our method improves registration accuracy and substantially reduces folding rates, indicating better deformation regularity. Although these lightweight baselines exhibit lower inference cost, their alignment quality remains inferior. Relative to recent coarse-to-fine correspondence-based approaches, our method achieves competitive Dice and folding performance on IXI and OASIS while requiring significantly less inference time, memory, and FLOPs. The advantage becomes more evident on IBSR18, which contains higher-resolution volumes. While the evaluated multi-scale baselines require input downsampling due to memory constraints, our method operates directly at full resolution and attains the best result among the evaluated IBSR18 configurations. These results demonstrate a favorable accuracy–efficiency trade-off and a fixed-memory scalability advantage, rather than universal superiority across all methods and protocols.

Figure 2 presents qualitative comparisons on IXI. Our method improves cortical alignment over VoxelMorph and TransMorph, particularly in sulcal and

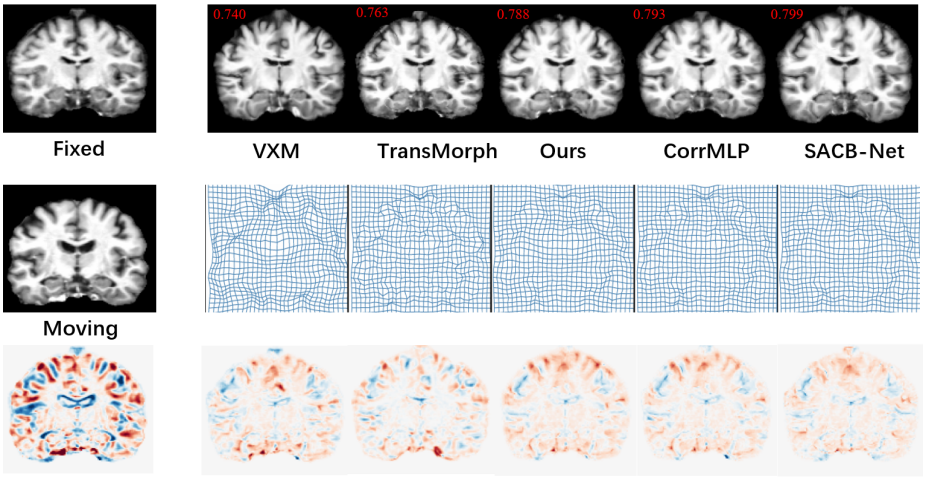

Fig. 2. Qualitative comparison on IXI. Top: warped images (Dice in red). Middle: deformation grids. Bottom: intensity difference maps between warped and fixed images.

ventricular regions, consistent with quantitative results. VoxelMorph exhibits irregular grid distortions, and TransMorph shows increased folding. Compared with CorrMLP and SACB-Net, our method further reduces residual misalignment along fine cortical structures, producing smoother deformation fields.

Ablation Study. We conduct experiments on IXI to assess each component. (1) *w/o Error Guidance:* Removing discrepancy-based routing reduces Dice and increases folding, highlighting the importance of semantic guidance. (2) *w/o Hard Non-Redundant Constraint:* Allowing repeated region selection slightly degrades accuracy and stability, indicating the benefit of non-redundant refinement. (3) *Dense Injection:* Replacing sparse Top-r routing with dense fusion yields similar Dice but significantly increases time and memory, demonstrating the efficiency advantage of structured sparsity. As shown in Table 2, the full model achieves the best overall trade-off between accuracy, stability, and efficiency.

Sensitivity to the Routing Budget r**.**We evaluate $r \in \{3, 6, 9\}$ with $N = 3$ fixed. Increasing r improves Dice with diminishing returns ($0.771 \rightarrow 0.785 \rightarrow 0.789$), while inference time, memory, and FLOPs grow approximately linearly. Overall, $r = 6$ achieves the lowest folding rate and a favorable balance among accuracy, efficiency, and smoothness; while $r = 9$ yields the best Dice score.

4 Conclusion

We present SPEAR, an error-guided sparse refinement framework for 3D deformable medical image registration. By selectively updating anatomically misaligned regions with non-redundant progressive refinement, SPEAR avoids dense multi-scale computation while maintaining strong alignment accuracy and deformation stability. The proposed design enables efficient high-resolution 3D

Table 2. Ablation and routing budget analysis on IXI ($N = 3$).

Ablation ($r = 6$)	Dice ↑	Foldings (%) ↓	Time (s) ↓	Memory (GB) ↓	FLOPs (G) ↓
w/o Error Guidance	0.729±0.152	0.214±0.089	0.341	25.6	390.16
w/o Hard NR	0.775±0.161	0.168±0.072	0.348	25.7	402.67
Dense Injection	0.781±0.158	0.140±0.066	0.612	33.9	810.52
$r = 3$	0.771±0.174	0.118±0.064	**0.312**	**23.8**	**366.92**
$r = 6$ (Selected)	0.785±0.169	**0.112±0.061**	0.349	25.7	423.58
$r = 9$	**0.789±0.165**	0.114±0.060	0.388	27.6	480.41

registration under limited GPU memory. Future work will extend the current evaluation beyond affine-prealigned inter-subject T1 brain MRI and investigate robustness under stronger domain shifts and non-linear deformations. In addition, more principled strategies are needed to mitigate potential biases introduced by frozen semantic priors from the auxiliary segmenter. We will further explore more statistically rigorous evaluation protocols on small-scale datasets such as IBSR18, and evaluate our method on larger, high-resolution datasets to further assess its robustness and generalizability.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62572308.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abulnaga, S.M., Hoopes, A., Dey, N., Hoffmann, M., Fischl, B., Guttag, J., Dalca, A.: Multimorph: On-demand atlas construction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 30906–30917 (2025)
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. Casamitjana, A., Sala-Llonch, R., Lekadir, K., Iglesias, J.E., Initiative, A.D.N., et al.: Uslr: An open-source tool for unbiased and smooth longitudinal registration of brain mri. Medical image analysis **105**, 103662 (2025)
4. 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)
5. Cheng, X., Zhang, T., Lu, W., Meng, Q., Frangi, A.F., Duan, J.: Sacb-net: Spatial-awareness convolutions for medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5227–5236 (2025)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 424–432. Springer (2016)

7. Consortium, I.: Ixi dataset (2003), available at https://brain-development.org/ixi-dataset/
8. Ding, Z., Niethammer, M.: Aladdin: Joint atlas building and diffeomorphic registration learning with pairwise alignment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20784–20793 (2022)
9. Fogarollo, S., Laimer, G., Bale, R., Harders, M.: Implicit deformable medical image registration with learnable kernels. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 249–259. Springer (2025)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (2015)
11. Lee, D., Alam, S., Jiang, J., Cervino, L., Hu, Y.C., Zhang, P.: Seq2morph: A deep learning deformable image registration algorithm for longitudinal imaging studies and adaptive radiotherapy. Medical physics **50**(2), 970–979 (2023)
12. Marcus, D.S., et al.: The open access series of imaging studies (oasis). Journal of Cognitive Neuroscience (2007)
13. Meng, L., Yang, J., Tian, R., Dai, X., Wu, Z., Gao, J., Jiang, Y.G.: Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. Advances in Neural Information Processing Systems **37**, 23464–23487 (2024)
14. Meng, M., Feng, D., Bi, L., Kim, J.: Correlation-aware coarse-to-fine mlps for deformable medical image registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9645–9654 (2024)
15. Rohlfing, T.: Internet brain segmentation repository (ibsr) (2012), available at https://www.nitrc.org/projects/ibsr
16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017)