RECURRENT AMORTIZED SOLVERS FOR DEFORMABLE IMAGE REGISTRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deformable image registration aligns anatomical structures across medical images, enabling comparison across subjects, modalities, and time points. It is commonly formulated as a regularized optimization problem balancing image alignment against deformation regularity. A central difficulty, shared by many regularized optimization problems, is choosing the regularization strength λ in advance: different choices induce different accuracy–regularity trade-offs. Current learning-based registration methods amortize this optimization across image pairs, but typically fix λ during training or use a single value throughout each prediction. We introduce the ***R**ecurrent **A**mortized **S**olver* (RAS), a meta solver that can condition unrolled optimization with flexible λ scheduling. RAS enables λ to be controlled throughout the unrolling process without retraining the model. On each of three public 3D image registration benchmarks, we show that our RAS solver reaches competitive accuracy–regularity fronts with a single trained model. Furthermore, we introduce a per-case Pareto framework that evaluates the trade-offs available to each image pair, extending registration evaluation beyond cohort averages. RAS therefore provides a general formulation for learning iterative solvers with inference-time regularization control, applicable to regularized optimization problems beyond image registration.

1 INTRODUCTION

Deformable image registration is among the most important tasks in medical image analysis. It estimates spatial correspondences between two anatomical images to enable direct comparison between subjects, modalities (e.g. CT and MRI), and time points. Good registration is defined by accurate anatomical alignment with a smooth and physiologically plausible deformation field (Sotiras et al., 2013; Chen et al., 2025). Classically, it is formulated as a *regularized optimization problem* (Modersitzki, 2009; Sotiras et al., 2013): given a pair of moving image I_m and a fixed image I_f, it estimates a dense spatial deformation $\phi : \Omega \to \Omega$ by solving

$$\phi^\star = \arg\min_{\phi} \mathcal{D}(I_m \circ \phi, I_f) + \lambda \mathcal{R}(\phi), \tag{1}$$

where Ω denotes the spatial domain, $I_m \circ \phi$ denotes the warped moving image by ϕ, $\mathcal{D}$ measures image dissimilarity, and $\mathcal{R}$ regularizes the estimated deformation. In practical implementations, λ is selected empirically per case (e.g. heart, lung), depending on the need for regularization (Klein et al., 2007; 2010; Reithmeir et al., 2024).

Recent deep learning methods have shifted medical image registration from instance-wise optimization (Eq. 1) to amortized learning. With learning-based methods, a neural network f_θ directly predicts deformation by mapping (I_m, I_f) to a parameterized deformation $\phi = f_\theta(I_m, I_f)$ (Balakrishnan et al., 2019; Chen et al., 2022; 2025). Using F_λ to denote the objective in Eq. 1 for a moving–fixed image pair, the network is trained over many pairs p_{train}:

$$\theta^\star = \arg\min_{\theta} \mathbb{E}_{(I_m, I_f)\sim p_{\text{train}}} \left[F_\lambda(f_\theta(I_m, I_f)) \right]. \tag{2}$$

Once trained, $f_{\theta^\star}(I_m, I_f)$ predicts $\widehat{\phi}$ for a new moving–fixed image pair. This process is called amortization, namely, investing computation in advance (over a large training dataset) and using it later without re-optimization, as exemplified by VoxelMorph and subsequent work (Amos, 2023;

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

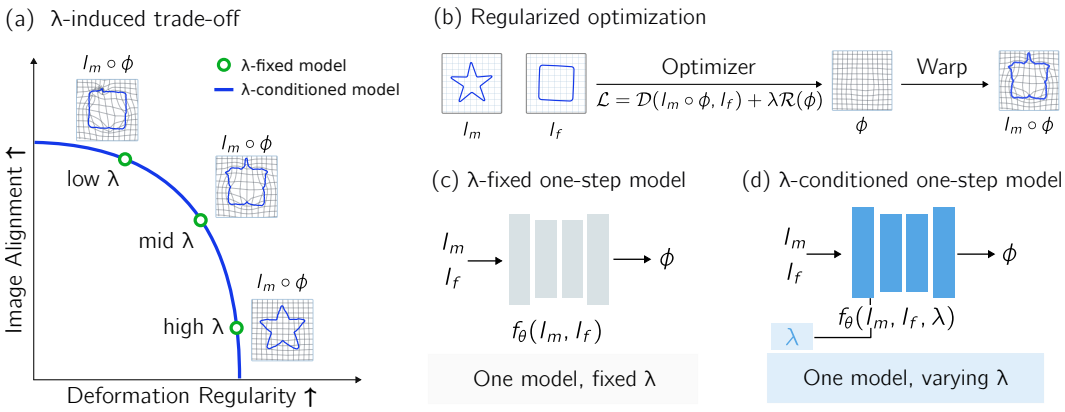

Figure 1: **Regularization trade-off in image registration.** (a) Schematic effect of λ on alignment and regularity. (b) Classical optimization estimates each image pair's deformation. One-step models predict (c) one trade-off at fixed λ or (d) different trade-offs conditioned on λ.

Balakrishnan et al., 2019; Chen et al., 2022; Tian et al., 2023; 2024). Amortized models yield much faster inference compared with classical optimization-based methods.

However, despite the drastic speed-up, amortized models have two intrinsic disadvantages. First, each model is trained with rigid regularization, i.e. a fixed λ. This makes it much more expensive to tweak λ than classical methods, since every λ now requires training a new model on all data. Second, the end-to-end mapping in Eq. 2 is essentially a one-step optimization. This deviates from the well-established principle of iterative optimization in classical methods, and makes the model difficult to generalize. The two issues hamper the reliable application of amortized models to medical image registration, as reliability is paramount for clinical applications.

Learning-based registration has evolved along two lines. The first is *λ-conditioned models* to address the first issue. These methods use one trained model to cover different λ's (Hoopes et al., 2021; Mok & Chung, 2021; 2022; Wang et al., 2023), i.e., they amortize not only over data but also over λ. This offers ease of tweaking λ, but is limited to end-to-end models. The second is *unrolling (recurrent) models* to address the second issue. Instead of end-to-end prediction (one-step optimization), these models improve the prediction quality over multiple learned steps (Gregor & LeCun, 2010; Monga et al., 2021; Yun et al., 2025; Zhao et al., 2019; Falta et al., 2022; Qiu et al., 2022; Zhang et al., 2025b;a). These methods enable iterative refinement, but models trained with a fixed regularization strength do not directly provide inference-time control of λ.

The two lines of research provide complementary yet incomplete solutions. Classical methods suggest that good registration demands both: iterative optimization and adaptive regularization (Allgower & Georg, 1993; Haber et al., 2000; Mang et al., 2019). Specifically, λ can be set larger at early steps to ensure global alignment, while smaller at later steps to allow local refinement (Haber et al., 2000; Himthani et al., 2022). This is analogous to the decaying step sizes in neural network training, which facilitates convergence and avoids bad local minima. Therefore, for reliable medical image registration, and for regularized optimization problems in general, it is desirable to enable both capabilities in one amortized solver.

We introduce the ***R**ecurrent **A**mortized **S**olver* (RAS), which represents such a meta-solver. The contributions of RAS are threefold:

1. A regularization-conditioned unrolling model for image registration. RAS combines regularization conditioning with unrolling in a single trained model. Each unrolling step can be conditioned on a specified λ_t at step t. This λ_t, called *λ-scheduling*, enables users to control the regularization throughout unrolling steps, without retraining the model.

2. Improved accuracy–regularity Pareto fronts. Varying regularization yields different operating points, whose nondominated solutions define an empirical Pareto front (Das & Dennis, 1998; Navon et al., 2021; Lin et al., 2022). RAS provides such a family of registration solutions by a single trained model. Across three 3D registration benchmarks, we demonstrate competitive empirical Pareto

fronts, with improved anatomical accuracy at comparable deformation regularity over parts of the evaluated range relative to competing models. Beyond varying a constant λ, per-update control of λ_t allows the regularization scheduling to shape these trade-offs during inference.

3. Per-case Pareto front for better evaluation. Current evaluation of image registration methods chooses one regularization setting for all image pairs based on average accuracy–regularity trade-offs (Sideri-Lampretsa et al., 2025; Grewal et al., 2024). But one regularization setting usually does not fit every case. We therefore introduce *per-case Pareto front*: RAS provides multiple candidate solutions for each image pair, bringing the accuracy–regularity trade-off to any individual case. Our label-free Per-case selection then uses deformation regularity alone to choose an output within the prescribed limit, whenever one is available. This selection rule applies to any λ-conditioned model; RAS additionally allows the candidate outputs to vary with the per-update scheduling. More broadly, this provides a generic framework for evaluating and deploying learned regularized solvers beyond cohort averages, with regularity controlled at the level of each individual case.

2 METHOD

2.1 BACKGROUND ON AMORTIZED IMAGE REGISTRATION

Let $z = (I_m, I_f)$ denote a moving–fixed image pair drawn from the training distribution p_{train}. We parameterize the spatial deformation by a displacement field $\phi : \Omega \to \mathbb{R}^3$, which maps a location $p \in \Omega$ to $p + \phi(p)$ in the three-dimensional space. Throughout the following formulation, we use $I_m \circ \phi$ to denote the warped moving image:

$$(I_m \circ \phi)(p) := I_m(p + \phi(p)) \,. \tag{3}$$

Under this parameterization, the registration objective in Eq. 1 is written as

$$F_{z,\lambda}(\phi) = \mathcal{D}(I_m \circ \phi, I_f) + \lambda \mathcal{R}(\phi), \tag{4}$$

where $\mathcal{D}$ measures image dissimilarity, $\mathcal{R}$ penalizes deformation irregularity, and $\lambda \geq 0$ controls their relative weight.

Amortized registration replaces repeated optimization for individual pairs with a predictor f_θ, whose parameters θ are learned from all training pairs. At a fixed λ, the predictor is trained using Eq. 2. At inference time, it directly produces the deformation $f_\theta(z)$ for a new moving–fixed image pair.

Regularization-conditioned registration extends this prediction to multiple regularization strengths within one model (Hoopes et al., 2021; Mok & Chung, 2021; 2022; Wang et al., 2023). During training, λ is sampled from a distribution p_λ:

$$\theta^\star_{\text{cond}} \in \arg\min_\theta \mathbb{E}_{\substack{z \sim p_{\text{train}} \\ \lambda \sim p_\lambda}} \left[F_{z,\lambda}(f_\theta(z;\lambda)) \right]. \tag{5}$$

This amortizes prediction across *both* image pairs and regularization strengths.

2.2 RECURRENT AMORTIZED SOLVER

We propose the recurrent amortized solver (RAS), an unrolled registration model that repeatedly applies a shared update network f_θ across unrolling steps (Gregor & LeCun, 2010; Monga et al., 2021). We first describe the recurrence at a single resolution. Starting from zero deformation $\phi_0 = 0$, RAS uses the current estimate ϕ_{t-1} to warp the moving image at step t, giving the network input

$$x_t = (I_m \circ \phi_{t-1}, I_f, \phi_{t-1}) \,. \tag{6}$$

The update network is additionally conditioned on the current regularization strength λ_t and a normalized step index τ_t:

$$q_t = (\lambda_t, \tau_t), \qquad \tau_t = \frac{t-1}{\max\{T-1, 1\}}, \quad t = 1, \ldots, T. \tag{7}$$

Here, λ_t specifies the regularization at the current step, while $\tau_t \in [0, 1]$ allows the shared network to distinguish early from late unrolling steps. Given x_t and q_t, the network predicts a residual deformation update

$$\Delta\phi_t = f_\theta(x_t; q_t), \tag{8}$$

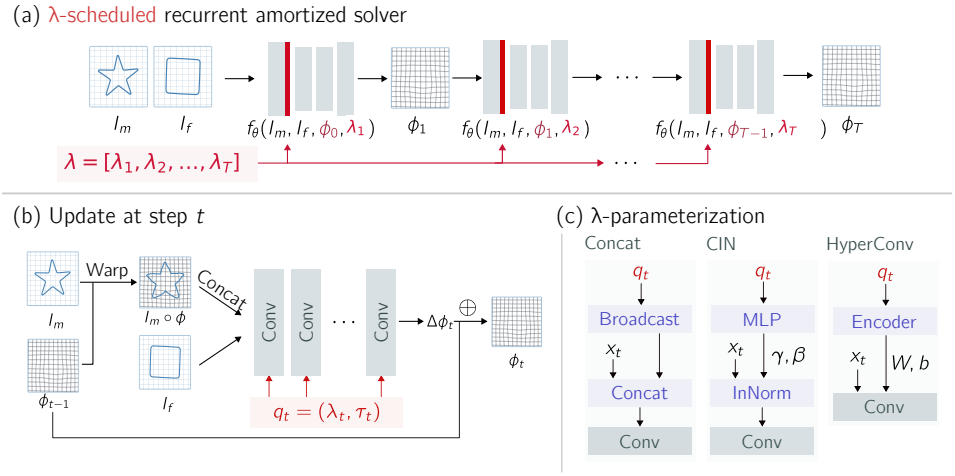

Figure 2: **Recurrent amortized solver (RAS).** (a) Shared recurrent updates follow a regularization schedule. (b) Each step warps the moving image and predicts a residual conditioned on $q_t = (\lambda_t, \tau_t)$, with normalized step index τ_t. (c) Concat, conditional instance normalization (CIN), and HyperConv provide alternative conditioning mechanisms.

and updates the current deformation as

$$\phi_t = \phi_{t-1} + \Delta\phi_t. \tag{9}$$

After each unrolling step, the original moving image is warped using ϕ_t, so the next step acts on the remaining misalignment. After T unrolling steps, ϕ_T is the final registration output.

2.3 λ-SCHEDULING AND CONDITIONING

The regularization strengths across the T unrolling steps form

$$\boldsymbol{\lambda} = \lambda_{1:T} = (\lambda_1, \ldots, \lambda_T), \tag{10}$$

which we call λ-*scheduling*. A constant λ-scheduling sets $\lambda_t = \bar{\lambda}$ for all t, while a non-constant λ-scheduling allows the regularization strength to change across unrolling steps. This follows the general idea of varying a problem parameter during an iterative solve (Allgower & Georg, 1993; Haber et al., 2000; Mang et al., 2019). In RAS, λ_t is provided directly to the shared update network at each unrolling step, allowing one trained model to use different choices of $\boldsymbol{\lambda}$ without changing θ.

When needed, we write $\phi_t^\theta(z; \boldsymbol{\lambda}, T)$ to make its dependence on the image pair z, the λ-scheduling, and the number of unrolling steps explicit. Appendix E.1 describes the multiresolution implementation.

We compare three ways of providing q_t to the shared update network.

Concatenation. The two scalars are broadcast over the spatial grid and concatenated with the registration state:

$$G_\theta^{\text{cat}}(x_t; q_t) = H_\theta\left([x_t, \lambda_t \mathbf{1}, \tau_t \mathbf{1}]\right). \tag{11}$$

Conditional instance normalization (CIN). An encoder E_θ maps q_t to an embedding $e_t = E_\theta(q_t)$. At feature layer r, CIN uses this embedding to modulate the normalized activations (Huang & Belongie, 2017; Dumoulin et al., 2016):

$$\text{CIN}_{\theta,r}(h_t^r; e_t) = (1 + \gamma_{\theta,r}(e_t)) \odot \text{IN}(h_t^r) + \beta_{\theta,r}(e_t). \tag{12}$$

HyperConv. Using the same embedding e_t, HyperConv generates the kernel and bias of convolutional layer r (Hoopes et al., 2021):

$$W_{\theta,r}(e_t) = \text{reshape}(A_r^W e_t), \qquad b_{\theta,r}(e_t) = A_r^b e_t, \tag{13}$$
$$h_t^{r+1} = \sigma\left(W_{\theta,r}(e_t) * h_t^r + b_{\theta,r}(e_t)\right).$$

2.4 LOSS FORMULATION

RAS is trained in a step-wise deep supervision manner, conditioned on the inputs. Specifically, let $\phi_t^\theta(z; \boldsymbol{\lambda}, T)$ denote intermediate output t generated by Eq. 9. Each state is supervised with its current objective $F_{z,\lambda_t}(\phi_t^\theta)$, evaluated on the input image grid. The training loss is

$$\mathcal{L}_{\mathrm{RAS}}(\theta) = \mathbb{E}_{\substack{z\sim p_{\mathrm{train}},\, T\sim p_T \\ \boldsymbol{\lambda}\sim p_{\boldsymbol{\lambda}}(\cdot|T)}} \left[\sum_{t=1}^{T} w_t \, F_{z,\lambda_t}\big(\phi_t^\theta(z;\boldsymbol{\lambda},T)\big) \right], \tag{14}$$

$$\theta^\star_{\mathrm{RAS}} \in \arg\min_\theta \mathcal{L}_{\mathrm{RAS}}(\theta), \qquad w_t = \frac{\kappa^{t-1}}{\sum_{s=1}^{T} \kappa^{s-1}}, \quad 0 < \kappa \le 1.$$

Primary training holds one sampled $\bar{\lambda}$ constant across the sequence. Appendix E.1 gives the multiresolution loss indexing and training protocol.

2.5 PARETO INFERENCE AND SOLUTION SELECTION

At inference, the network parameters θ are fixed. By changing the inference setting $a = (\boldsymbol{\lambda}, T)$, the same RAS checkpoint produces multiple candidate registrations for each image pair. These outputs provide different accuracy–regularity trade-offs without retraining the model.

We compare two ways of selecting from these outputs under a deformation irregularity limit b.

Shared selection. Standard model selection chooses one setting for the whole cohort based on average performance. We call this *Shared*: among settings whose mean deformation irregularity is below b, it selects the one with the highest mean anatomical accuracy. Because the limit is imposed only on the cohort average, individual registrations may still exceed it.

Per-case selection. We instead select an output separately for each image pair. Let $Q_i(a)$ denote the measured deformation irregularity of case i under setting a. Among the outputs satisfying the limit, *Per-case* selects the one closest to the limit:

$$a_i^{\mathrm{pc}}(b) \in \arg\max_{a:\, Q_i(a)\le b} Q_i(a). \tag{15}$$

This rule uses only the predicted deformation field and therefore does not require anatomical labels.

For cohort-level evaluation, we plot mean anatomical accuracy against mean deformation irregularity across the evaluated settings; the nondominated outputs form the empirical Pareto front. At the individual level, the nondominated candidate outputs for each image pair define its *per-case Pareto front*.

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We evaluate OASIS (Marcus et al., 2007), NLST (Hansen et al., 2025), and LPBA40 (Shattuck et al., 2008), with 314/20/80 train/validation/test subjects, 170/10/30 pairs, and 35/5 train/evaluation subjects, respectively. Each dataset is trained separately. Anatomical accuracy uses Dice on OASIS/LPBA40 and landmark TRE on NLST. Regularity uses NDV, the percentage of voxels with nonpositive Jacobian determinant, and SDLogJ, the spatial variation of log Jacobian determinants.

The main comparison crosses VoxelMorph (Balakrishnan et al., 2019), single-pass LapIRN (Mok & Chung, 2020), and RAS with Concat, CIN, and HyperConv conditioning. RAS uses a LapIRN-based backbone. All models use NCC and diffusion losses; NLST additionally uses shared anatomical supervision. Constant-λ sweeps use the same grid within each benchmark. We also include GraDIRN (Qiu et al., 2022), RIIR (Zhang et al., 2025b), CorrMLP (Meng et al., 2024), and TransMorph (Chen et al., 2022) as OASIS references. Appendices A.3 and E give model costs, update counts, checkpoint selection, grids, and uncertainty estimates.

3.2 ACCURACY–REGULARITY TRADE-OFFS ACROSS BENCHMARKS

In the main comparison against VoxelMorph and single-pass LapIRN, RAS has the strongest Shared accuracy across all three datasets under each conditioning method, both without an NDV limit and at NDV $\leq$ .3%. Conditioning changes both accuracy and the regularity at which it is attained (Table 1). On OASIS, Concat and CIN RAS both round to .772 Dice at their best outputs, but their achieved NDV is .326% and .233%, respectively. Concat RAS attains the lowest TRE on NLST and the highest Dice on LPBA40. The detailed Pareto support, with reference learning-based methods is shown in Fig. 3. Further analysis on NLST and LPBA can be found in Appendix A.

Table 1: **Shared accuracy across three 3D benchmarks, mean (SD).** Each method selects the cohort-wide λ maximizing evaluation accuracy, either unconstrained (Best) or at mean NDV $\leq$ .3%. These are cohort oracle selections. Bold/underline mark the best/second-best backbone within each conditioning group.

	OASIS Dice ↑		NLST TRE ↓		LPBA40 Dice ↑	
Method	Best	NDV $\leq$ .3%	Best	NDV $\leq$ .3%	Best	NDV $\leq$ .3%
Concatenation						
VoxelMorph	.692(.051)	.691(.053)	4.26(2.74)	4.26(2.74)	.658(.032)	.658(.032)
LapIRN	.769(.026)	.768(.027)	2.73(1.88)	2.81(2.00)	.710(.013)	.710(.013)
RAS	**.772(.026)**	**.772(.027)**	**1.38(.98)**	**1.39(1.04)**	**.719(.010)**	**.719(.010)**
Conditional instance normalization						
VoxelMorph	.702(.050)	.701(.052)	4.59(4.05)	4.59(4.05)	.658(.032)	.658(.032)
LapIRN	.768(.027)	.766(.028)	3.76(3.42)	3.76(3.42)	.711(.011)	.711(.011)
RAS	**.772(.025)**	**.772(.025)**	**2.01(1.57)**	**2.01(1.57)**	**.715(.010)**	**.715(.010)**
HyperConv						
VoxelMorph	.692(.051)	.690(.052)	4.53(2.71)	4.54(2.77)	.653(.032)	.653(.032)
LapIRN	.746(.030)	.744(.032)	4.02(2.45)	4.08(2.65)	.688(.017)	.688(.017)
RAS	**.761(.026)**	**.761(.026)**	**2.76(1.71)**	**2.76(1.71)**	**.706(.011)**	**.706(.011)**

On OASIS, Concat RAS contributes 17 of the 20 joint front points of RAS and single-pass LapIRN, and achieves higher interval-average Dice over their common NDV interval (Fig. 3). CIN and HyperConv parameterizations show the same positive ordering. Figure 4 shows selected anatomical and landmark alignments across the three benchmarks.

Paired RAS–LapIRN accuracy differences and uncertainty are reported in Appendix A.2; a matched recurrent comparison further supports near-specialist quality under constant regularization (Appendix A.4).

3.3 REGULARIZATION STRENGTH AND UPDATE COUNT

On OASIS, more unrolling steps improve Dice at comparable regularity, with gains largely saturating after five steps (Fig. 5). At the finest level, one to three updates contribute to the joint Dice–NDV front; further updates improve final NCC–diffusion without Dice gains. On NLST, fewer coarse-level updates lower landmark accuracy at comparable regularity even with the finest-level depth fixed (Appendix B.1).

3.4 BENEFITS OF PER-UPDATE REGULARIZATION CONTROL

With the checkpoint and 18-update budget fixed, a validation-selected schedule uses $\lambda = 1/16$ at the two coarser levels and $\lambda = 1/2$ at the finest level. It lowers mean NLST test TRE by .065–.132 mm across three seeds relative to the best of 449 constant-λ settings constrained by the schedule's achieved test mean NDV (Appendix B.3, including paired uncertainty).

For within-level control, each library uses 129 queries of 18 updates. Free schedules vary six finest-level strengths, holding coarser levels at the final strength. Averaging best-feasible-accuracy differences over a common NDV interval, these schedules improve OASIS validation Dice by .022–.080 percentage points across three seeds. Every evaluated family improves NLST validation TRE across all three seeds, whereas OASIS hold–release remains negative (Appendix B.6).

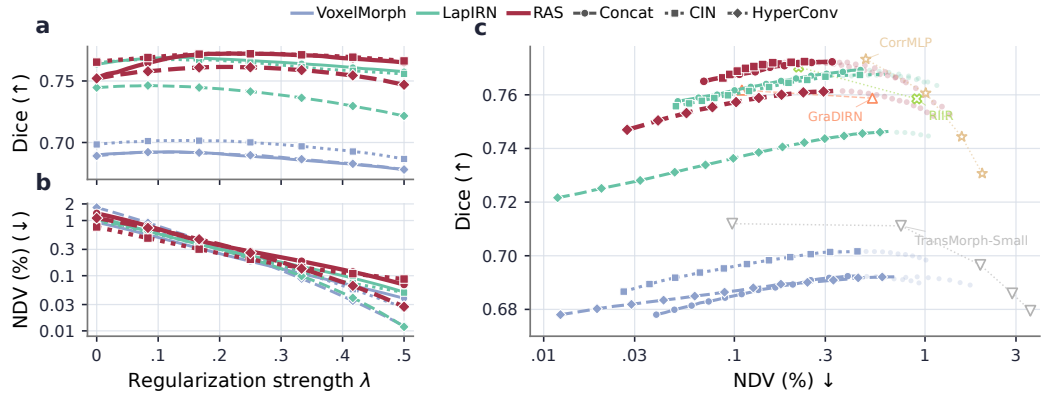

Figure 3: λ **responses (a,b) and empirical Pareto fronts (c)** across 31 sampled λ values. In (c), RAS dominates competing sampled outputs over part of the range: it achieves no lower Dice and no higher NDV, with a strict improvement in at least one. Opaque curves join within-model nondominated outputs; faint points are dominated within the same model; hollow markers denote reference methods.

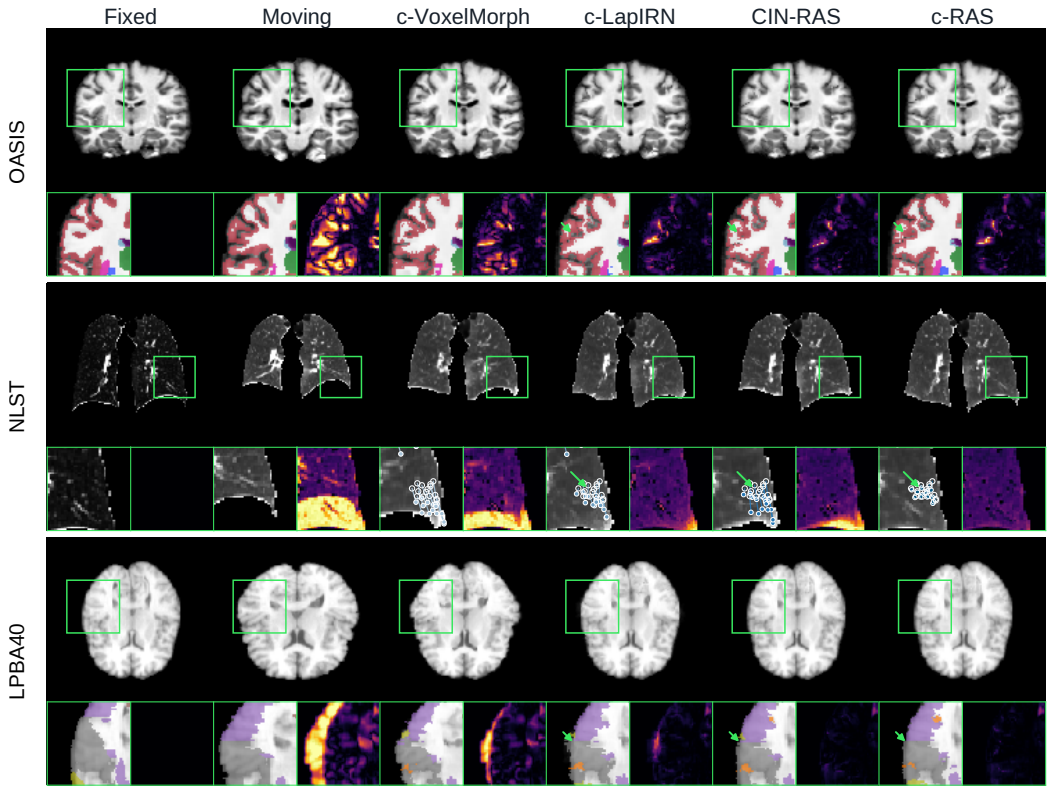

Figure 4: **Registration examples on OASIS, NLST, and LPBA40** (c: concatenation). Green boxes mark crops; lower panels show label or landmark alignment and absolute intensity differences.

With the checkpoint, strengths, and update count fixed, increasing rather than reversed order reduces NDV and SDLogJ on all 80 OASIS test pairs at nine settings, with small Dice losses at some settings. Interior timing also changes regularity on both datasets with the finest-level endpoints fixed (Appendix B.2). Scheduling-aware fine-tuning further improves image matching on both datasets while changing regularity; anatomical effects vary across operating ranges (Appendix B.4, Fig. 10).

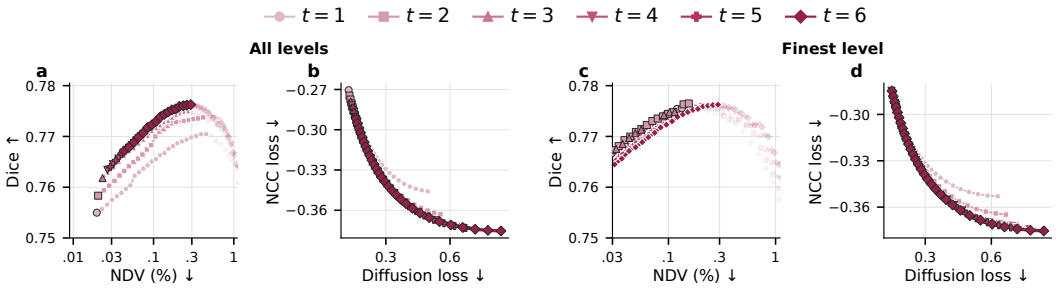

Figure 5: **Unrolling-step sensitivity on OASIS.** All panels use one frozen checkpoint, 20 validation pairs, and 31 λ queries. (a,b) Steps per level: (t, t, t); (c,d): $(6, 6, t)$. Both NCC and diffusion losses are computed on each configuration's final output. Dark marker edges identify outputs nondominated across the six step counts. Panel (c) starts at NDV 0.03%.

3.5 Per-case selection under individual regularity limits

Per-case selection uses measured NDV to meet individual limits without anatomical labels. Figure 6 shows that a Shared setting can satisfy the average limit while individual pairs exceed it. At the lowest NDV limit tested for all OASIS methods, Shared Concat-RAS exceeds the limit for 34 of 80 pairs; Per-case satisfies it for all 80.

Across methods and NDV limits, Per-case brings all Shared violations within the limit, with a median Dice change of -0.18 percentage points. This requires a feasible candidate; selecting by NDV alone does not ensure the highest Dice or lowest TRE (Appendix E.2).

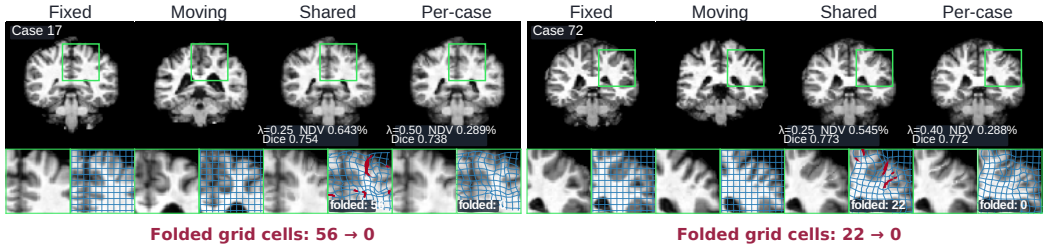

Figure 6: **Per-case selection reduces folding while meeting the individual NDV limit.** Shared achieves a cohort-mean NDV of 0.275%, below the 0.289% limit, but exceeds it for the two largest-violation pairs shown. Per-case meets the limit for both. Red marks folded in-plane grid cells, whose counts fall from 56 to 0 and from 22 to 0; these counts describe the displayed grids, while NDV is measured over the full deformation.

4 Discussion

RAS combines recurrent refinement with per-update regularization control. At fixed checkpoints and update budgets, scheduling recovers outputs beyond dense constant-strength sweeps, while order reversal shows that timing changes the finite solve. These candidate solutions support Per-case selection under individual regularity limits.

Refinement improves accuracy (Fig. 5) at additional cost: Concat RAS takes 157 ms per 128^3 OASIS pair and setting on one H100, versus 39 ms for LapIRN and 7 ms for VoxelMorph. Each additional setting adds inference cost. With architecture and 18 updates matched, conditioned RAS retains near-specialist quality (Appendix A.4). Our exploratory external-refinement extension also improves frozen LapIRN's accuracy–regularity trade-off (Appendix B.5).

This control extends beyond registration: within one knee-MRI checkpoint, a decreasing schedule improves PSNR and SSIM over constant regularization while increasing total variation (Ap-

pendix D.1). Local objective sensitivity and a finite-horizon perturbation bound complement this empirical extension, motivating regularization control in other inverse problems (Appendix D.2).

Limitations. The main architecture–conditioning matrix uses one seed per configuration; separate three-seed experiments assess RAS scheduling on validation data. Case bootstrap measures cohort uncertainty. NLST scheduling reuses a previously analyzed test cohort, two seed-specific TRE intervals include zero, and matched scheduling fine-tuning does not consistently improve anatomical accuracy. NDV feasibility alone does not establish anatomical correctness. Jointly improving image matching and regularity remains an open question.

5 Related Work

Conditional and iterative registration. Conditional registration allows users to vary λ after training (Hoopes et al., 2021; Mok & Chung, 2021; 2022; Wang et al., 2023). Iterative methods refine deformations through repeated learned or optimization-based updates (Zhao et al., 2019; Falta et al., 2022; Qiu et al., 2022; Zhang et al., 2025b;a; Jena et al., 2025; 2024a). Related approaches condition updates on the iteration index (Yun et al., 2025) or adapt the update sequence through latent policy optimization (Zhang et al., 2026). RAS combines regularization conditioning with recurrent unrolling, allowing users to specify λ_t separately at each update of one trained model.

Amortized optimization and learned solvers. Amortized optimization learns approximate solution mappings over a family of problems (Amos, 2023). Learning to optimize learns reusable update rules, providing a meta-learning perspective on iterative solvers (Andrychowicz et al., 2016). RAS adopts this perspective for $F_{z,\lambda}$ by learning a shared recurrent update rule across image pairs and regularization strengths, using a weighted sum of intermediate objective values for supervision. Continuation methods motivate varying problem parameters during optimization (Allgower & Georg, 1993). RAS makes regularization strength an explicit input to this learned update rule, allowing users to control λ_t throughout inference.

Registration evaluation and individual regularity. Learn2Reg evaluates accuracy, robustness, deformation plausibility, and runtime (Hering et al., 2023; Hansen et al., 2025), while Jena et al. examine how supervision and domain shift affect comparisons between learned and classical methods (Jena et al., 2024b). Accuracy–regularity curves further characterize trade-offs across registration settings (Sideri-Lampretsa et al., 2025). We introduce a per-case Pareto framework to evaluate each image pair's candidate solution set, making individual trade-offs explicit beyond cohort averages. Our label-free Per-case rule selects from this candidate set using measured deformation regularity alone, meeting the prescribed individual limit whenever a feasible output is available.

6 Conclusion

RAS unifies algorithm unrolling and step-wise regularization control in a single trained solver, making the strength and timing of regularization part of learned iterative optimization. Our experiments show that schedules shape the final deformation and can extend empirical accuracy–regularity fronts beyond constant-λ inference. The per-case Pareto framework further extends evaluation beyond cohort averages, characterizing the trade-offs available to each image pair. More broadly, RAS provides a methodological framework for regularized optimization beyond image registration, allowing learned iterative solvers to adapt their solution trade-offs through inference-time scheduling without retraining.

AI Use Statement

Generative AI tools assisted with manuscript drafting and language editing, literature search, experimental design, implementation and debugging of experimental and analysis code, interpretation of results, figure preparation, and the formulation, derivation, and writing of mathematical claims and proofs. The authors directed these activities and take responsibility for the final manuscript, including all methods, mathematical claims, experimental results, and conclusions.

REFERENCES

Eugene L. Allgower and Kurt Georg. Continuation and path following. *Acta Numerica*, 2:1–64, 1993. doi: 10.1017/S0962492900002336. URL https://www.cambridge.org/core/journals/acta-numerica/article/continuation-and-path-following/4368C662C0FA6F729FA4B2A5C1B60085.

Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023. doi: 10.1561/2200000102.

Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/fb87582825f9d28a8d42c5e5e5e8b23d-Abstract.html.

Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John Guttag, and Adrian V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019. doi: 10.1109/TMI.2019.2897538.

J. Frédéric Bonnans and Alexander Shapiro. *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research. Springer, New York, 2000. doi: 10.1007/978-1-4612-1394-9.

Junyu Chen, Erin C. Frey, Yufan He, W. Paul Segars, Ye Li, and Yong Du. TransMorph: Transformer for unsupervised medical image registration. *Medical Image Analysis*, 82:102615, 2022. doi: 10.1016/j.media.2022.102615.

Junyu Chen, Yihao Liu, Shuwen Wei, Zhangxing Bian, Shalini Subramanian, Aaron Carass, Jerry L. Prince, and Yong Du. A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond. *Medical Image Analysis*, 100:103385, 2025. doi: 10.1016/j.media.2024.103385.

Indraneel Das and John E. Dennis. Normal-boundary intersection: A new method for generating the pareto surface in nonlinear multicriteria optimization problems. *SIAM Journal on Optimization*, 8(3):631–657, 1998. doi: 10.1137/S1052623496307510.

Asen L. Dontchev and R. Tyrrell Rockafellar. Robinson's implicit function theorem and its extensions. *Mathematical Programming*, 117(1–2):129–147, 2009. doi: 10.1007/s10107-007-0161-1.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016.

Fenja Falta, Lasse Hansen, and Mattias P. Heinrich. Learning iterative optimisation for deformable image registration of lung CT with recurrent convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13436 of *Lecture Notes in Computer Science*, pp. 301–309. Springer, 2022. doi: 10.1007/978-3-031-16446-0_29.

Anthony V. Fiacco. Sensitivity analysis for nonlinear programming using penalty methods. *Mathematical Programming*, 10(1):287–311, 1976. doi: 10.1007/BF01580677.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 399–406, 2010. URL https://icml.cc/Conferences/2010/papers/449.pdf.

Monika Grewal, Henrike Westerveld, Peter Bosman, and Tanja Alderliesten. Multi-objective learning for deformable image registration. In *Proceedings of the 7th International Conference on Medical Imaging with Deep Learning*, volume 250 of *Proceedings of Machine Learning Research*, pp. 538–553. PMLR, 2024. URL https://proceedings.mlr.press/v250/grewal24a.html.

Eldad Haber, Uri M. Ascher, and Doug Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16(5):1263–1280, 2000. doi: 10.1088/0266-5611/16/5/309.

Lasse Hansen, Wiebke Heyer, Christoph Großbröhmer, Frederic Madesta, Thilo Sentker, Wang Jiazheng, Yuxi Zhang, Hang Zhang, Min Liu, Junyi Wang, et al. Learn2reg 2024: New benchmark datasets driving progress on new challenges. *arXiv preprint arXiv:2509.01217*, 2025.

Alessa Hering, Lasse Hansen, Tony C. W. Mok, Albert C. S. Chung, Hanna Siebert, Stephanie Hager, Annkristin Lange, Sven Kuckertz, Stefan Heldmann, Wei Shao, Sulaiman Vesal, Mirabela Rusu, Geoffrey Sonn, Theo Estienne, Maria Vakalopoulou, Luyi Han, Yunzhi Huang, Pew-Thian Yap, Mikael Brudfors, Yael Balbastre, Samuel Joutard, Marc Modat, Gal Lifshitz, Dan Raviv, Jinxin Lv, Qiang Li, Vincent Jaouen, Dimitris Visvikis, Constance Fourcade, Mathieu Rubeaux, Wentao Pan, Zhe Xu, Bailiang Jian, Francesca De Benetti, Marek Wodzinski, Niklas Gunnarsson, Jens Sjolund, Daniel Grzech, Huaqi Qiu, Zeju Li, Alexander Thorley, Jinming Duan, Christoph Grosbrohmer, Andrew Hoopes, Ingerid Reinertsen, Yiming Xiao, Bennett Landman, Yuankai Huo, Keelin Murphy, Nikolas Lessmann, Bram van Ginneken, Adrian V. Dalca, and Mattias P. Heinrich. Learn2Reg: Comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning. *IEEE Transactions on Medical Imaging*, 42(3):697–712, 2023. doi: 10.1109/TMI.2022.3213983. URL https://doi.org/10.1109/TMI.2022.3213983.

Naveen Himthani, Malte Brunn, Jae-Youn Kim, Miriam Schulte, Andreas Mang, and George Biros. CLAIRE—parallelized diffeomorphic image registration for large-scale biomedical imaging applications. *Journal of Imaging*, 8(9):251, 2022. doi: 10.3390/jimaging8090251.

Andrew Hoopes, Malte Hoffmann, Bruce Fischl, John V. Guttag, and Adrian V. Dalca. HyperMorph: Amortized hyperparameter learning for image registration. In *Information Processing in Medical Imaging*, volume 12729 of *Lecture Notes in Computer Science*, pp. 3–17. Springer, 2021. doi: 10.1007/978-3-030-78191-0_1.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 1510–1519. IEEE, 2017.

Rohit Jena, Pratik Chaudhari, and James C. Gee. FireANTs: Adaptive riemannian optimization for multi-scale diffeomorphic matching, 2024a. URL https://arxiv.org/abs/2404.01249.

Rohit Jena, Deeksha Sethi, Pratik Chaudhari, and James C. Gee. Deep learning in medical image registration: Magic or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pp. 108331–108353, 2024b. doi: 10.52202/079017-3439. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c3fe2a07ec47b89c50e89706d2e23358-Abstract-Conference.html.

Rohit Jena, Pratik Chaudhari, and James C. Gee. Deep implicit optimization enables robust learnable features for deformable image registration. *Medical Image Analysis*, 103:103577, 2025. doi: 10.1016/j.media.2025.103577. URL https://doi.org/10.1016/j.media.2025.103577.

Stefan Klein, Marius Staring, and Josien P. W. Pluim. Evaluation of optimization methods for nonrigid medical image registration using mutual information and B-splines. *IEEE Transactions on Image Processing*, 16(12):2879–2890, 2007. doi: 10.1109/TIP.2007.909412.

Stefan Klein, Marius Staring, Keelin Murphy, Max A. Viergever, and Josien P. W. Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010. doi: 10.1109/TMI.2009.2035616.

Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, and Qingfu Zhang. Pareto set learning for expensive multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19231–19247, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7a583691ccfcf8945ab714b677ccbf0b-Abstract.html.

Andreas Mang, Amir Gholami, Christos Davatzikos, and George Biros. CLAIRE: A distributed-memory solver for constrained large deformation diffeomorphic image registration. *SIAM Journal on Scientific Computing*, 41(5):C548–C584, 2019. doi: 10.1137/18M1207818.

Daniel S. Marcus, Tracy H. Wang, Jamie Parker, John G. Csernansky, John C. Morris, and Randy L. Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9): 1498–1507, 2007. doi: 10.1162/jocn.2007.19.9.1498. URL https://doi.org/10.1162/jocn.2007.19.9.1498.

Mingyuan Meng, Dagan Feng, Lei Bi, and Jinman Kim. Correlation-aware coarse-to-fine mlps for deformable medical image registration. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9645–9654. IEEE, 2024.

Jan Modersitzki. *FAIR: Flexible Algorithms for Image Registration*. SIAM, Philadelphia, 2009. doi: 10.1137/1.9780898718843.

Tony C. W. Mok and Albert C. S. Chung. Large deformation diffeomorphic image registration with laplacian pyramid networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12263 of *Lecture Notes in Computer Science*, pp. 211–221. Springer, 2020. doi: 10.1007/978-3-030-59716-0_21.

Tony C. W. Mok and Albert C. S. Chung. Conditional deformable image registration with convolutional neural network. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12904 of *Lecture Notes in Computer Science*, pp. 35–45. Springer, 2021. doi: 10.1007/978-3-030-87202-1_4.

Tony C. W. Mok and Albert C. S. Chung. Conditional deep laplacian pyramid image registration network in Learn2Reg challenge. In *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis – MICCAI 2021 Challenges*, volume 13166 of *Lecture Notes in Computer Science*. Springer, 2022. doi: 10.1007/978-3-030-97281-3_23.

Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. doi: 10.1109/MSP.2020.3016905. URL https://arxiv.org/abs/1912.10557.

Aviv Navon, Aviv Shamsian, Ethan Fetaya, and Gal Chechik. Learning the Pareto front with hypernetworks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=NjF772F4ZZR.

Huaqi Qiu, Kerstin Hammernik, Chen Qin, Chen Chen, and Daniel Rueckert. Embedding gradient-based optimization in image registration networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, volume 13436 of *Lecture Notes in Computer Science*, pp. 56–65. Springer, 2022. doi: 10.1007/978-3-031-16446-0_6.

Anna Reithmeir, Julia A. Schnabel, and Veronika A. Zimmer. Learning physics-inspired regularization for medical image registration with hypernetworks. In *Medical Imaging 2024: Image Processing*, volume 12926 of *Proc. SPIE*, pp. 129262K, 2024. doi: 10.1117/12.3006539.

Stephen M. Robinson. Strongly regular generalized equations. *Mathematics of Operations Research*, 5(1):43–62, 1980. doi: 10.1287/moor.5.1.43.

Torsten Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *IEEE Transactions on Medical Imaging*, 31(2):153–163, 2012. doi: 10.1109/TMI.2011.2163944.

David W. Shattuck, Mubeena Mirza, Vitria Adisetiyo, Cornelius Hojatkashani, Georges Salamon, Katherine L. Narr, Russell A. Poldrack, Robert M. Bilder, and Arthur W. Toga. Construction of a 3d probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064–1080, 2008. doi: 10.1016/j.neuroimage.2007.09.031. URL https://doi.org/10.1016/j.neuroimage.2007.09.031.

Vasiliki Sideri-Lampretsa, Daniel Rueckert, and Huaqi Qiu. Evaluation of deformable image registration under alignment-regularity trade-off, 2025. URL https://arxiv.org/abs/2503.07185.

Aristeidis Sotiras, Christos Davatzikos, and Nikos Paragios. Deformable medical image registration: A survey. *IEEE Transactions on Medical Imaging*, 32(7):1153–1190, 2013. doi: 10.1109/TMI.2013.2265603.

Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, and Marc Niethammer. GradICON: Approximate diffeomorphisms via gradient inverse consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18084–18094, 2023.

Lin Tian, Hastings Greer, Roland Kwitt, François-Xavier Vialard, Raúl San José Estépar, Sylvain Bouix, Richard Rushmore, and Marc Niethammer. uniGradICON: A foundation model for medical image registration. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15002 of *Lecture Notes in Computer Science*, pp. 749–760. Springer, 2024. doi: 10.1007/978-3-031-72069-7_70.

Yinsong Wang, Huaqi Qiu, and Chen Qin. Conditional deformable image registration with spatially-variant and adaptive regularization. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ISBI53787.2023.10230464.

Junno Yun, Yaşar Utku Alçalar, and Mehmet Akçakaya. Time-embedded algorithm unrolling for computational MRI. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/86f19768c110b25ac41dc08cd3783d12-Abstract-Conference.html.

Jure Zbontar, Florian Knoll, Anuroop Sriram, Tullie Murrell, Zhengnan Huang, Matthew J. Muckley, Aaron Defazio, Ruben Stern, Patricia Johnson, Mary Bruno, Marc Parente, Krzysztof J. Geras, Joe Katsnelson, Hersh Chandarana, Zizhao Zhang, Michal Drozdzal, Adriana Romero, Michael Rabbat, Pascal Vincent, Nafissa Yakubova, James Pinkerton, Duo Wang, Erich Owens, C. Lawrence Zitnick, Michael P. Recht, Daniel K. Sodickson, and Yvonne W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018. URL https://arxiv.org/abs/1811.08839.

Runxun Zhang, Yizhou Liu, Dongrui Li, Bo Xu, and Jingwei Wei. MorphSeek: Fine-grained latent representation-level policy optimization for deformable image registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Yi Zhang, Yidong Zhao, and Qian Tao. Bridging classical and learning-based iterative registration through deep equilibrium models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15962 of *Lecture Notes in Computer Science*, pp. 88–98. Springer, 2025a. doi: 10.1007/978-3-032-04947-6_9.

Yi Zhang, Yidong Zhao, Hui Xue, Peter Kellman, Stefan Klein, and Qian Tao. Recurrent inference machine for medical image registration. *Medical Image Analysis*, 106:103748, 2025b. doi: 10.1016/j.media.2025.103748.

Shengyu Zhao, Yue Dong, Eric I-Chao Chang, and Yan Xu. Recursive cascaded networks for unsupervised medical image registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10600–10610, 2019. doi: 10.1109/ICCV.2019.01070.

A Additional Model Comparisons

A.1 Accuracy and achieved regularity

Tables 2–4 give the selected λ and achieved regularity for every accuracy entry in Table 1. Best and NDV-limited selection use each method's own cohort-wide setting.

Table 2: **OASIS: regularity at the Table 1 operating points.** NDV and SDLogJ are mean (SD) over 80 test pairs. An equals sign repeats the corresponding Best entry.

		Best			NDV $\leq .3\%$		
Backbone	Conditioning	λ	NDV (%)	SDLogJ	λ	NDV (%)	SDLogJ
VoxelMorph	Concat	.1167	.441 (.161)	1.399 (.230)	.1833	.281 (.122)	1.124 (.216)
VoxelMorph	CIN	.1500	.442 (.134)	1.408 (.196)	.2167	.283 (.103)	1.135 (.186)
VoxelMorph	HyperConv	.1167	.690 (.190)	1.742 (.220)	.2167	.283 (.108)	1.131 (.191)
LapIRN	Concat	.1333	.459 (.116)	1.440 (.169)	.2167	.277 (.087)	1.132 (.160)
LapIRN	CIN	.0833	.614 (.132)	1.658 (.167)	.2167	.269 (.084)	1.118 (.155)
LapIRN	HyperConv	.0833	.643 (.120)	1.698 (.149)	.2000	.298 (.073)	1.177 (.130)
RAS	Concat	.2167	.326 (.095)	1.229 (.163)	.2500	.275 (.087)	1.135 (.160)
RAS	CIN	.2167	.233 (.073)	1.056 (.145)	=	=	=
RAS	HyperConv	.2167	.330 (.076)	1.243 (.131)	.2333	.294 (.071)	1.177 (.129)

Table 3: **NLST: regularity at the Table 1 operating points.** NDV and SDLogJ are mean (SD) over 30 test pairs. An equals sign repeats the corresponding Best entry.

		Best			NDV $\leq .3\%$		
Backbone	Conditioning	λ	NDV (%)	SDLogJ	λ	NDV (%)	SDLogJ
VoxelMorph	Concat	.0000	.024 (.033)	.375 (.177)	=	=	=
VoxelMorph	CIN	.0000	.072 (.104)	.543 (.301)	=	=	=
VoxelMorph	HyperConv	.0000	.514 (.690)	1.264 (.829)	.0833	.272 (.389)	.938 (.619)
LapIRN	Concat	.1667	.488 (.547)	1.314 (.687)	.4167	.254 (.370)	.905 (.603)
LapIRN	CIN	.0000	.134 (.128)	.763 (.305)	=	=	=
LapIRN	HyperConv	.0000	.611 (.550)	1.555 (.583)	.1667	.221 (.273)	.921 (.461)
RAS	Concat	.1667	.423 (.445)	1.243 (.630)	.3333	.280 (.311)	1.011 (.542)
RAS	CIN	.0000	.100 (.091)	.684 (.271)	=	=	=
RAS	HyperConv	.0833	.159 (.138)	.838 (.312)	=	=	=

NLST CIN-LapIRN reports the retrained seed-1 endpoint; Appendix E details its supervision and optimization.

Table 4: **LPBA40: regularity at the Table 1 operating points.** NDV and SDLogJ are mean (SD) over 20 ordered pairs from five held-out subjects. Best also satisfies NDV $\leq .3\%$ for every model.

Backbone	Conditioning	λ	NDV (%)	SDLogJ
VoxelMorph	Concat	.2333	.115 (.153)	.649 (.379)
VoxelMorph	CIN	.1667	.180 (.162)	.863 (.340)
VoxelMorph	HyperConv	.3167	.063 (.078)	.510 (.275)
LapIRN	Concat	.1833	.149 (.056)	.836 (.147)
LapIRN	CIN	.1500	.171 (.060)	.891 (.148)
LapIRN	HyperConv	.1167	.259 (.101)	1.083 (.198)
RAS	Concat	.2167	.163 (.073)	.866 (.181)
RAS	CIN	.2333	.058 (.025)	.554 (.105)
RAS	HyperConv	.1500	.176 (.074)	.903 (.177)

A.2 PAIRED ACCURACY DIFFERENCES

Table 5 compares RAS and LapIRN on matched evaluation pairs. Each of 5,000 paired bootstrap resamples reselects each method's cohort-wide λ under the reported rule. The intervals quantify uncertainty in cohort oracle performance.

Table 5: **Paired RAS–LapIRN accuracy gains for Table 1.** Positive values favor RAS. OASIS: Dice percentage points over 80 pairs; NLST: TRE reduction in mm over 30 pairs. Brackets give 95% paired bootstrap intervals with λ reselected in every resample. For NLST Concat under the NDV limit, intervals use the 4,641 of 5,000 resamples in which both methods have a feasible query; all other comparisons use all 5,000.

Dataset	Conditioning	Best	NDV $\leq .3\%$
OASIS	Concat	$+0.305$ $[0.151, 0.481]$	$+0.469$ $[0.273, 0.653]$
	CIN	$+0.475$ $[0.333, 0.612]$	$+0.632$ $[0.415, 0.815]$
	HyperConv	$+1.508$ $[1.274, 1.760]$	$+1.756$ $[1.481, 2.111]$
NLST	Concat	$+1.348$ $[0.971, 1.863]$	$+1.415$ $[0.991, 1.926]$
	CIN	$+1.753$ $[1.088, 2.609]$	$+1.753$ $[1.088, 2.609]$
	HyperConv	$+1.261$ $[0.892, 1.732]$	$+1.326$ $[0.900, 1.899]$

A.3 MODEL SIZE AND INFERENCE COST

HyperConv narrows LapIRN/RAS channels to match Concat parameter counts within 0.4%; VoxelMorph retains the same widths. Table 6 reports their inference costs.

Table 6: **Concat and HyperConv inference cost on OASIS.** One 128^3 pair on a single H100. Time is median [IQR] over 50 runs; peak is allocated CUDA memory.

Model	Parameters	Steps	Time (ms)	Peak (MiB)
VoxelMorph (Concat)	57,091	1	7.03 [0.02]	751.0
VoxelMorph (HyperConv)	908,576	1	7.18 [0.03]	777.8
LapIRN (Concat)	894,441	3	39.37 [1.04]	3248.4
LapIRN (HyperConv)	897,728	3	27.76 [0.32]	943.5
RAS (Concat)	900,057	18	156.67 [1.17]	3608.4
RAS (HyperConv)	897,744	18	95.99 [3.92]	1303.5

A.4 COMPARISON WITH FIXED-λ SPECIALISTS

One conditioned RAS model closely follows seven fixed-λ recurrent specialists on OASIS, with mean matched-front gaps of .000985 Dice under NDV and .001071 under SDLogJ. Architecture, channels, time conditioning, training protocol, 18 updates, and warp count are matched. Specialists retain four of the 13 joint front points, quantifying trade-offs not reached by the shared model. Table 7 instead compares outputs at identical λ.

Table 7: **Fixed-λ specialists and conditioned RAS on OASIS.** Each cell lists specialist / conditioned-model mean (SD) over the same 80 test pairs.

λ	Dice $\uparrow$	NDV (%) $\downarrow$	SDLogJ $\downarrow$
0	.699 (.028) / .753 (.026)	12.635 (1.500) / 1.349 (.220)	6.909 (.343) / 2.431 (.189)
$1/12$	.766 (.026) / .765 (.025)	.885 (.167) / .795 (.158)	1.983 (.178) / 1.885 (.177)
$2/12$	.773 (.026) / .771 (.026)	.498 (.121) / .433 (.112)	1.503 (.170) / 1.408 (.169)
$3/12$	.773 (.027) / .772 (.027)	.311 (.089) / .275 (.088)	1.200 (.156) / 1.135 (.161)
$4/12$	.772 (.029) / .771 (.028)	.170 (.062) / .183 (.070)	.906 (.142) / .936 (.152)
$5/12$	.768 (.030) / .768 (.028)	.090 (.041) / .117 (.054)	.681 (.125) / .761 (.141)
$6/12$	.765 (.031) / .765 (.030)	.040 (.024) / .069 (.039)	.486 (.099) / .602 (.124)

A.5 Responses and fronts on additional benchmarks

Figures 7 and 8 extend Fig. 3 to NLST and LPBA40. They show both the response to λ and accuracy at achieved NDV, using the same colors and panel layout.

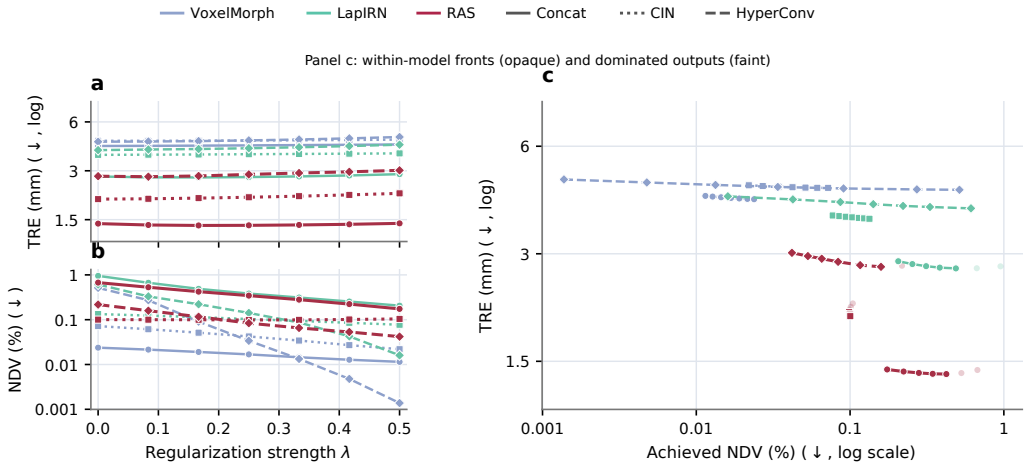

Figure 7: **NLST responses (a,b) and fronts (c).** Cohort means over 30 test pairs at seven λ values per model. Opaque curves in (c) join within-model nondominated outputs; faint points show dominated outputs. CIN-LapIRN uses the retrained seed-1 endpoint and the same seven-value evaluation grid as Table 1.

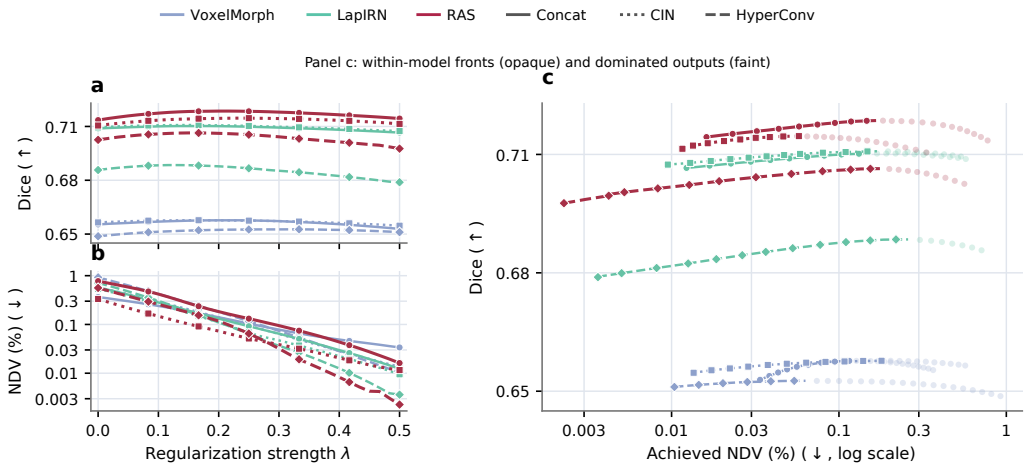

Figure 8: **LPBA40 responses (a,b) and fronts (c).** Cohort means over 20 ordered pairs from five held-out subjects at 31 λ values per model, using the same models as Table 1. Opaque curves in (c) join within-model nondominated outputs; faint points show dominated outputs.

B Regularization and Update Controls

B.1 Training update count and supervision

Table 8 tests whether sampling shorter training sequences preserves the accuracy–regularity front. Varying the update count at every level reduces the average training sequence from 18 to 10.5 updates, with closely aligned fronts at 18 inference updates. Varying only the finest level also reduces inference to eight updates, so this comparison includes a change in inference depth. The table also compares intermediate and final-output supervision.

Table 8: **Training ablations on 20 OASIS validation pairs.** Peak Dice is mean (SD). ΔDice is the average difference from the first row over common regularity ranges. Other variants supervise intermediate outputs; the final-output variant keeps variable updates at every level.

Training configuration	Mean train updates	Inference updates	Peak Dice	ΔDice at NDV	ΔDice at SDLogJ
Variable updates at every level	10.5	18	.778 (.019)	+0.0000	+0.0000
Six updates at every level	18.0	18	.779 (.019)	-0.0004	-0.0004
Variable updates at the finest level	5.5	8	.776 (.019)	-0.0011	-0.0010
Supervision only at the final output	10.5	18	.780 (.019)	+0.0016	+0.0016

Coarse-resolution updates. On 10 NLST validation pairs, keeping six finest-level updates and reducing the two coarser levels from six updates each to one, two, or three raises TRE at matched NDV by 1.309, .756, or .370 mm under constant regularization. The corresponding changes under hold–release are 1.307, .748, and .356 mm. Coarse refinement therefore contributes accuracy even when the finest-level work is unchanged.

B.2 Sensitivity to update order

A frozen Concat-RAS uses 18 updates, with anchor λ at both coarser levels and six evenly spaced finest-level strengths from $\lambda - a\min(\lambda, .5 - \lambda)$ to $\lambda + a\min(\lambda, .5 - \lambda)$, applied in increasing or reversed order. The nine settings combine $\lambda \in \{2/15, 1/4, 11/30\}$ and $a \in \{1/3, 2/3, 1\}$. On 20 OASIS validation pairs, increasing order lowers NDV and SDLogJ for every pair at every setting, without significant Dice changes after Holm correction.

On all 80 OASIS test pairs, the same settings again reduce both regularity metrics for every pair. Mean NDV reductions span .0256–.4849 percentage points; mean Dice differences span $-.00174$ to $+.00006$, with significant decreases at six settings after Holm correction (Fig. 9, Table 9).

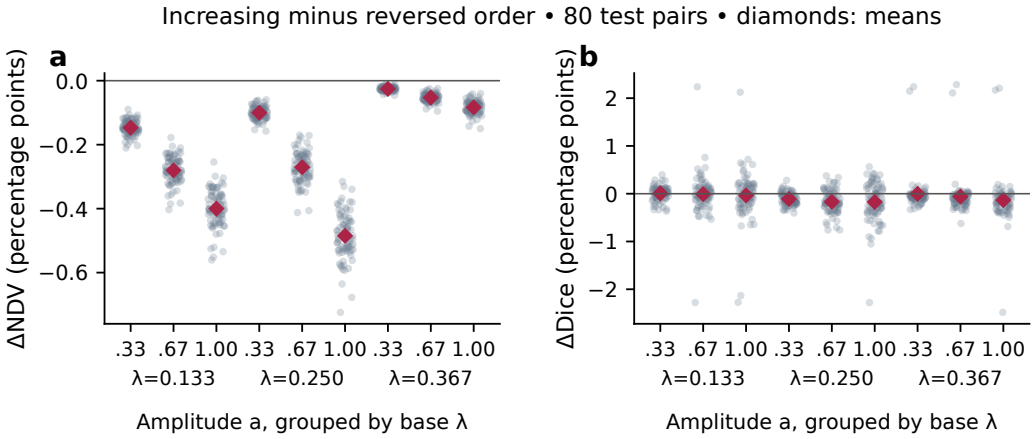

Figure 9: **Update-order effects on 80 OASIS test pairs.** Each dot is one pair's difference between increasing and reversed order for the same six finest-level strengths. Diamonds denote means. All NDV differences are negative; Dice changes vary with the anchor and amplitude.

Fixed-endpoint control. To isolate intermediate timing, we hold both coarser levels at q and place one zero at finest-level step 2, 3, 4, or 5, with q at every other update. For $q \in \{.25, .36328125, .5\}$, these schedules share all level endpoints, the complete strength multiset, the checkpoint, and 18 updates. On 20 OASIS validation pairs (two seeds) and 10 NLST validation pairs (three seeds), mean NDV differs between every pair of positions: 36 contrasts on OASIS and 54 on NLST, with absolute differences of .053–.337 and .000075–.0203 percentage points, respectively. These descriptive

Table 9: **Test-set update-order differences.** Increasing minus reversed order, mean (SD) across 80 pairs. Dice uses two-sided paired Wilcoxon tests with Holm correction over nine settings. NDV is in percentage points; Dice is on the $[0, 1]$ scale.

λ	a	ΔDice	Adjusted p	ΔNDV (pp)	ΔSDLogJ
0.133	0.33	+0.0001 (0.0015)	0.835	-0.1471 (0.0235)	-0.1985 (0.0148)
0.133	0.67	-0.0001 (0.0043)	1.000	-0.2804 (0.0425)	-0.3639 (0.0245)
0.133	1.00	-0.0004 (0.0051)	1.000	-0.4000 (0.0583)	-0.4969 (0.0327)
0.250	0.33	-0.0011 (0.0013)	$< .001$	-0.1004 (0.0200)	-0.1871 (0.0165)
0.250	0.67	-0.0017 (0.0024)	$< .001$	-0.2705 (0.0487)	-0.4500 (0.0350)
0.250	1.00	-0.0017 (0.0042)	0.004	-0.4849 (0.0752)	-0.6947 (0.0467)
0.367	0.33	-0.0000 (0.0037)	$< .001$	-0.0256 (0.0064)	-0.0648 (0.0077)
0.367	0.67	-0.0006 (0.0038)	$< .001$	-0.0528 (0.0131)	-0.1318 (0.0152)
0.367	1.00	-0.0013 (0.0048)	$< .001$	-0.0834 (0.0208)	-0.2030 (0.0228)

contrasts reuse the same cases and show that intermediate controls affect the final deformation even when the endpoints are fixed.

B.3 Inference with a validation-selected schedule on NLST

Three independently trained Concat-RAS models use the same validation-selected schedule: $\lambda = 1/16$ at the two coarser levels and $1/2$ at the finest, with six updates per level. For each seed, the constant reference minimizes mean TRE without exceeding the scheduled output's mean NDV. Its 449 settings comprise 257 values in $[0, .5]$, 128 in $(.5, 1]$, and 64 in $(1, 2]$, including strengths above the training range.

Table 10: **Validation-selected scheduling on 30 NLST test pairs.** TRE is mean (SD); reductions favor scheduling. Intervals use 20,000 paired bootstrap resamples, reselecting the feasible constant reference in each. Both settings share the checkpoint and 18 updates.

Seed	NDV limit (%)	Constant TRE	Scheduled TRE	Reduction [95% CI] (mm)
0	.085	1.76 (1.85)	1.62 (1.40)	.132 [.014,.335]
1	.074	1.90 (1.82)	1.83 (1.61)	.066 [-.013,.197]
2	.187	1.75 (1.29)	1.69 (.98)	.065 [-.022,.220]

All three mean reductions are positive, but two intervals include zero.

B.4 Training with a regularization schedule

Hold–release keeps both coarser levels at $\bar{\lambda}$ and uses $(.5, .5, .5, .5, \sqrt{.5\bar{\lambda}}, \bar{\lambda})$ at the finest. On 20 OASIS validation pairs, inference-only scheduling (C/R) lowers interval-average Dice at matched NDV by .00076 relative to C/C. Scheduling-aware training (R/R) recovers .00076 relative to C/R, without a significant gain at any of four NDV limits. This is not evidence of an advantage over constant fine-tuning.

On 30 NLST test pairs, scheduling and constant controls both receive 50 fine-tuning epochs and use 18 inference updates. Scheduling-aware training changes TRE by $+.01459$ mm over the common NDV range; at NDV $\leq .20\%$, its TRE reduction is .01136 mm [95% paired bootstrap CI: $-.00119, .03033$]. Anatomical gains are therefore inconsistent. Relative to C/R, R/R lowers mean final NCC loss on both datasets at identical queried λ values while increasing diffusion (Fig. 10), improving image matching while changing regularity.

Auxiliary anatomical supervision. A single-seed pilot on 20 OASIS development pairs matches initialization, training budget, and 449 constant plus 240 nonconstant queries. Dice-loss weights $0/.1/1$ yield constant-regularization Dice $.7736/.8022/.8017$ and scheduling gains of $.069/.018/.012$ Dice percentage points at matched NDV ($.057/.014/.009$ at SDLogJ), all averaged over common regularity ranges. Anatomical supervision raises accuracy but reduces the additional scheduling gain; this pilot does not explain cross-dataset differences.

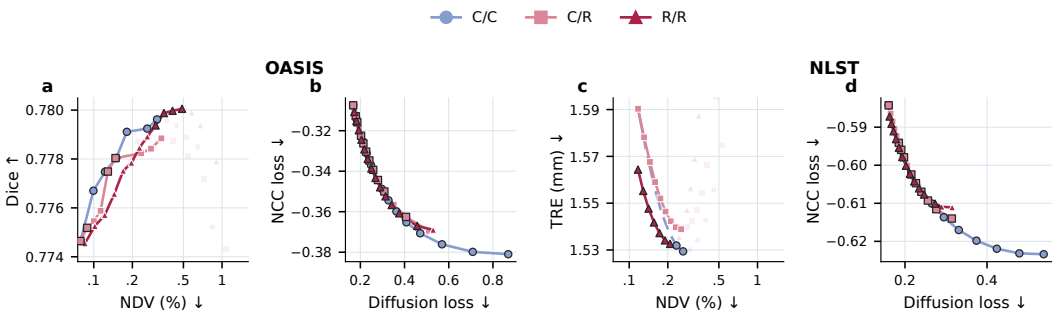

Figure 10: **Scheduling fine-tuning on OASIS and NLST.** C: constant; R: hold–release; settings denote training/inference. Lines join within-setting nondominated outputs; faint points show all evaluated outputs. Dark marker edges identify outputs nondominated across the three settings in each panel. Intensity panels use final-output NCC loss and diffusion; NLST uses the matched constant-finetuning control. Panel (c) displays TRE 1.525–1.595 mm.

B.5 Exploratory extension: external refinement of LapIRN

To explore iterative refinement beyond the standard single-pass baseline, we construct an inference-time extension of frozen, pretrained Concat-LapIRN. This extension adds an outer residual-refinement loop without retraining, whereas RAS learns its shared updates through unrolled training and exposes regularization control within that process. We evaluate $K \in \{1, 2, 4, 6\}$ passes on 80 OASIS test pairs at 31 strengths in $[0, .5]$. Each pass predicts a residual v_{k+1} from the fixed image and the original moving image warped by the accumulated displacement u_k, starting from $u_0 = 0$. Unlike the additive RAS updates, residual pullbacks compose as $u_{k+1}(p) = v_{k+1}(p) + u_k(p + v_{k+1}(p))$, with frozen weights and one strength throughout. One pass exactly reproduces the original Dice, NDV, and SDLogJ values. Each pass invokes three resolution levels; $3K$ calls are not FLOP-matched to RAS updates.

At identical λ, two passes lower cohort-mean NDV at all 31 settings and improve Dice at 23. They approach RAS's peak Dice with lower NDV (Table 11). Over common NDV support [.0692,.8000]%, their interval-average feasible Dice envelope exceeds RAS by .00060. Four and six passes lower peak accuracy. These results show that external refinement can improve a pretrained conditional model. This experiment holds λ constant across passes; the fixed-checkpoint RAS experiments in Section 3.4 separately assess the effect of varying regularization within the learned recurrence.

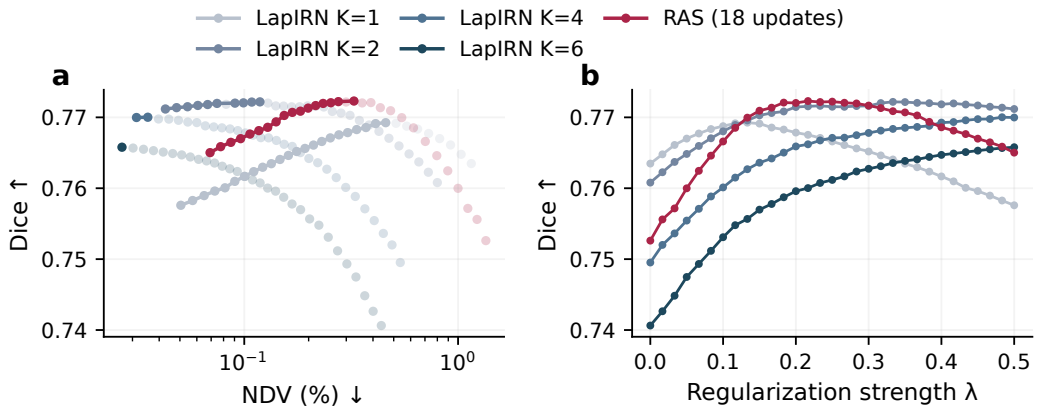

Figure 11: **External refinement of frozen LapIRN, constructed in this work.** Cohort means over 80 OASIS test pairs and 31 strengths. (a) Lines join within-method strictly nondominated outputs; faint points show all outputs. (b) Dice responses to regularization strength. RAS uses 18 recurrent updates; LapIRN uses K complete three-level passes.

Table 11: **Exploratory external-refinement extension.** $K = 1$ is standard LapIRN; $K > 1$ adds the outer refinement loop constructed here. Each method selects its own peak-Dice λ; NDV is measured at that setting. The constrained column maximizes cohort-mean Dice over settings with mean NDV at most .3%. These are cohort oracle selections.

Method	Peak Dice	NDV (%)	Dice at NDV $\leq$ .3%
LapIRN, $K = 1$	0.7692	0.4588	0.7675
LapIRN, $K = 2$	0.7722	0.1184	0.7722
LapIRN, $K = 4$	0.7700	0.0354	0.7700
LapIRN, $K = 6$	0.7658	0.0268	0.7658
RAS (18 updates)	0.7723	0.3259	0.7722

B.6 Independent training-seed repeatability

RAS seeds 0, 1, and 2 are trained independently from scratch. OASIS uses fixed 50-epoch EMA checkpoints and 20 validation pairs; NLST uses fixed raw 100-epoch checkpoints and 10 validation pairs. Every query uses 18 updates. These runs measure training variation separately from case-bootstrap uncertainty.

Each library contains 129 queries in $[0, .5]$: uniformly spaced strengths for the constant reference, or nine shared constants and 120 schedules. Per-level schedules assign one strength per resolution. Smooth and free schedules vary the finest-level strengths through three interpolated knots or six separately specified values, respectively, while holding the coarser levels at the final strength. Hold–release uses the constant-grid endpoints.

Table 12 compares the best feasible accuracy of each scheduling library against constant regularization. Gains are averaged over the same NDV interval within each seed, with equal query and update budgets; the interval can differ between seeds.

Table 12: **Scheduling-family gains across three training seeds.** Equal 129-query budgets; 18 updates per query. Entries are validation-envelope Dice gains (OASIS, percentage points) or TRE reductions (NLST, mm), not test effects or confidence intervals.

Dataset	Scheduling family	Seed 0	Seed 1	Seed 2
OASIS	Hold–release, finest level	-0.0311	-0.0536	-0.0312
	Per-level constant	+0.0470	+0.0248	+0.0004
	Smooth, finest level	+0.0132	+0.0357	+0.0079
	Free, finest level	+0.0801	+0.0340	+0.0224
NLST	Hold–release, finest level	+0.0130	+0.0328	+0.0304
	Per-level constant	+0.0309	+0.0626	+0.0297
	Smooth, finest level	+0.0125	+0.0290	+0.0257
	Free, finest level	+0.0119	+0.0325	+0.0211

Gains depend on the scheduling family and NDV budget; OASIS hold–release remains negative across all three seeds.

C Additional Output-Selection Results

Figures 12 and 13 extend the OASIS selection comparison to NLST and LPBA40. Population and Oracle use anatomical accuracy; Per-case uses only deformation regularity.

D Extensions and Sensitivity Analysis

D.1 Knee MRI reconstruction

Multicoil knee MRI uses data consistency plus $\lambda\mathrm{TV}$, with ten fastMRI training volumes and 20 disjoint evaluation volumes under a common forward model and training budget (Zbontar et al., 2018; Yun et al., 2025). Table 13 compares a constant-path-trained Amortized model with a separate mixed-path-trained RAS model. Constant, increasing, and decreasing inference use the same RAS

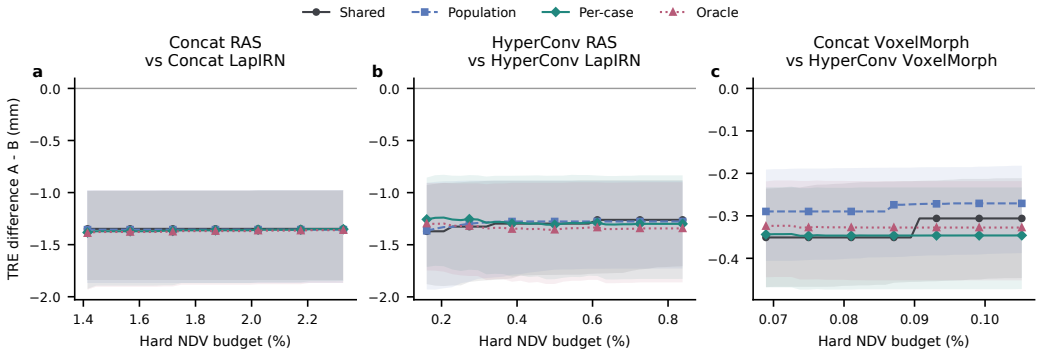

Figure 12: **NLST TRE differences across selection rules.** Comparisons use common feasible NDV intervals; shading shows paired-case bootstrap uncertainty.

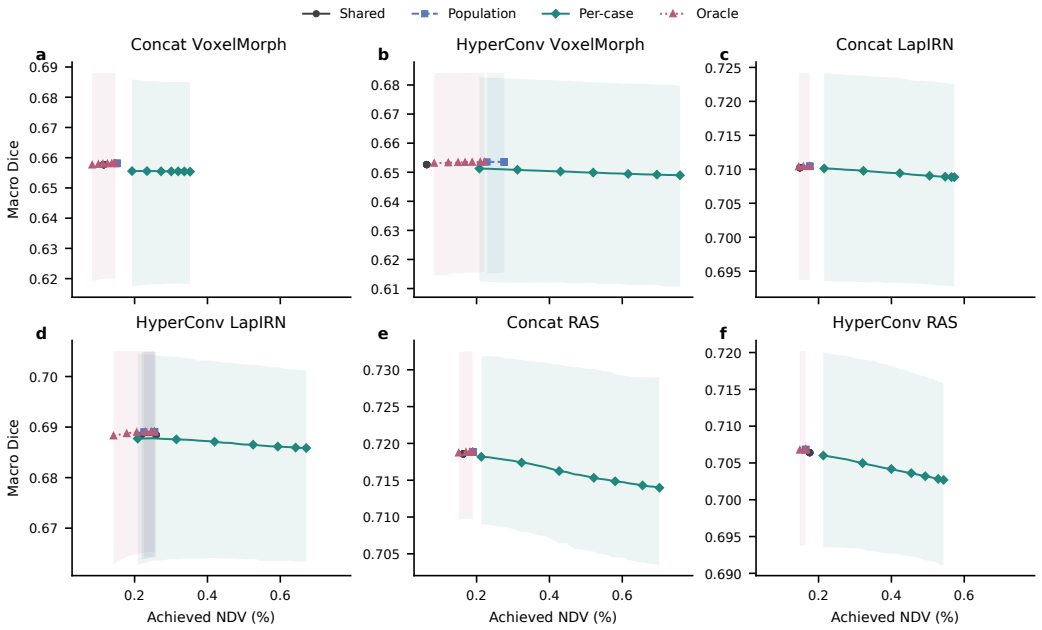

Figure 13: **LPBA40 accuracy across selection rules.** Rules use identical evaluated outputs; shading shows held-out-subject bootstrap uncertainty.

checkpoint and five updates. Metrics are averaged over five λ values within each volume before paired comparison across volumes.

Table 13: **Regularization strength and update order in knee MRI.** Mean improvements [95% bootstrap CI] over 20 paired volumes; positive values favor the first method. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ after Holm correction.

Comparison	ΔPSNR (dB)	ΔSSIM	ΔD ($\times 10^{-4}$)	ΔTV ($\times 10^{-5}$)
Amortized − fixed	-0.077^{***} $[-0.099, -0.059]$	-0.0018^{***} $[-0.0023, -0.0013]$	-0.25 $[-0.53, 0.12]$	169.99^{***} $[48.43, 321.00]$
RAS constant − amortized	-0.044^{**} $[-0.068, -0.020]$	-0.0003 $[-0.0008, 0.0001]$	0.37^{***} $[0.30, 0.43]$	11.22 $[-1.49, 27.85]$
Increasing − RAS constant	-0.012^{***} $[-0.017, -0.007]$	-0.0012^{***} $[-0.0014, -0.0010]$	-0.67^{***} $[-0.72, -0.62]$	2.37^{***} $[2.26, 2.47]$
Decreasing − RAS constant	0.036^{***} $[0.030, 0.043]$	0.0028^{***} $[0.0025, 0.0031]$	1.29^{***} $[1.18, 1.41]$	-8.20^{***} $[-8.84, -7.62]$

D.2 MATHEMATICAL SENSITIVITY

Objective sensitivity. Let $F_z(\phi, \lambda) = \mathcal{D}_z(\phi) + \lambda\mathcal{R}(\phi)$ be twice continuously differentiable near a stationary point with positive-definite Hessian. The implicit function theorem gives a locally unique differentiable branch of strict local minimizers $\phi_z^\star(\lambda)$ with

$$\frac{\partial \phi_z^\star}{\partial \lambda} = -[\nabla_\phi^2 F_z(\phi_z^\star, \lambda)]^{-1} \nabla_\phi \mathcal{R}(\phi_z^\star). \tag{16}$$

If the Hessian stays bounded below by μI, $\mu > 0$, and $\|\nabla\mathcal{R}\| \le G$ along the branch, integration gives $\|\phi_z^\star(\lambda_1) - \phi_z^\star(\lambda_2)\| \le (G/\mu)|\lambda_1 - \lambda_2|$ (Fiacco, 1976; Robinson, 1980; Bonnans & Shapiro, 2000; Dontchev & Rockafellar, 2009). This local structure motivates sharing a conditional solver across nearby regularization strengths, whose solutions vary smoothly along the same nondegenerate branch. Objective sensitivity does not determine anatomical accuracy (Rohlfing, 2012).

Finite control dependence. For a fixed per-resolution unrolling-step vector $\mathbf{T}$, let $T_\Sigma = \sum_\ell T_\ell$ denote the total number of unrolling steps and write $\phi_t = \Phi_{\theta,t}(\phi_{t-1}, z, \lambda_t)$ for $t = 1, \dots, T_\Sigma$, where $\Phi_{\theta,t}$ includes the fixed step coordinate and resolution-level transition.

Proposition 1 (Finite-horizon control stability)**.** *Suppose that, on the reachable state set,* $\|\Phi_{\theta,t}(\phi, z, a) - \Phi_{\theta,t}(\psi, z, a)\| \le K_t^\phi \|\phi - \psi\|$ *and* $\|\Phi_{\theta,t}(\phi, z, a) - \Phi_{\theta,t}(\phi, z, b)\| \le K_t^\lambda |a - b|$. *Then two schedules with the same initialization satisfy*

$$\left\| \phi_{T_\Sigma}^\theta(z; \boldsymbol{\lambda}, \mathbf{T}) - \phi_{T_\Sigma}^\theta(z; \widetilde{\boldsymbol{\lambda}}, \mathbf{T}) \right\| \le \sum_{t=1}^{T_\Sigma} K_t^\lambda |\lambda_t - \widetilde{\lambda}_t| \prod_{s=t+1}^{T_\Sigma} K_s^\phi. \tag{17}$$

Proof. Let $e_t = \|\phi_t - \widetilde{\phi}_t\|$. The two Lipschitz conditions and the triangle inequality give $e_t \le K_t^\phi e_{t-1} + K_t^\lambda |\lambda_t - \widetilde{\lambda}_t|$. Starting from $e_0 = 0$ and unrolling this recurrence yields Eq. 17. □

This bound controls perturbation propagation under the stated assumptions; the existence and direction of an order effect are assessed experimentally.

E REPRODUCIBILITY

E.1 IMPLEMENTATION AND TRAINING

The multiresolution implementation proceeds from coarse to fine, sharing update parameters within each level and using separate parameters across levels. The displacement at one level initializes the next after resampling, and the step coordinate resets at each level. We write $\mathbf{T} = (T_1, \dots, T_L)$ for the update allocation and $T_\Sigma = \sum_{\ell=1}^L T_\ell$ for its total. The recurrent-only model has $S = T_\Sigma$ supervised outputs. In Eq. 14, the multiresolution loss sums over these chronological outputs with weights $w_t = \kappa^{t-1} / \sum_{s=1}^S \kappa^{s-1}$; intermediate displacements are resampled to the input image grid.

Models are trained separately for each dataset using Adam, local NCC, and diffusion regularization. All NLST models additionally use the same final-output three-class mask Dice loss (weight .1), including keypoint regions, without a direct coordinate loss. Main comparisons use 50 training epochs on OASIS, 15 on LPBA40, and 90–140 on NLST depending on the model. Each model uses one fixed checkpoint across all queried λ values; the main OASIS comparison uses exponential moving averages. LPBA40 evaluates all 20 ordered pairs from five held-out subjects, without a separate validation/test split.

Primary RAS training samples one $\bar{\lambda} \sim \mathrm{Unif}[0, .5]$ per sequence and holds it constant across updates. Prefix lengths $T_\ell \in \{1, \dots, 6\}$ are sampled independently per level. Inference uses six recurrent updates per level. Loss weights follow Eq. 14 with $\kappa = .8$; fixed-λ specialists retain the same recurrent architecture and training protocol. Scheduling fine-tuning and matched constant controls instead emphasize later outputs with $w_t \propto .8^{S-t}$.

NLST CIN-LapIRN is retrained with a lower learning rate, gradient clipping, and supervision at all three resolutions, using a separate seed. Its protocol is developed on 10 validation pairs and evaluated on 30 test pairs. This is a stabilized baseline with a modified training protocol.

Base initializers. The main NLST Concat/HyperConv RAS models include an additional base prediction before the recurrent updates at each resolution. It predicts a displacement increment from the current registration state; its output and conditioning strength are included in the chronological training sequence, giving $S = T_\Sigma + L$. The scheduling experiments use the recurrent-only model with 18 updates.

E.2 Evaluation and selection

Main comparisons use 31 strengths $\{k/60\}_{k=0}^{30}$ on OASIS/LPBA40 and seven on NLST, with splits in Section 3.1. Shared selects one cohort-wide setting using anatomical accuracy; Per-case uses only deformation regularity and requires a feasible candidate. Population maximizes $U_i(a) - \eta Q_i(a)$, with U_i denoting Dice or negative TRE and a shared η chosen to meet the cohort budget. Oracle selects the most accurate feasible output per pair. Population and Oracle require anatomical labels. Appendix B.4 specifies hold–release and auxiliary-supervision query budgets.

E.3 Metrics and uncertainty

Dice averages 35 OASIS or 56 LPBA40 structures; NLST uses keypoint TRE in mm. NDV is the percentage of voxels with nonpositive Jacobian determinant. SDLogJ is the spatial standard deviation of $\log(\mathrm{clip}(J, 10^{-9}, 10^{9}))$, averaged over pairs. Case-level metrics are reported as mean (SD), paired effects with 95% confidence intervals where available, and latency as median [IQR].

Front comparisons use paired resampling and common regularity support. OASIS front intervals use 10,000 bootstrap resamples; registration scheduling intervals use 20,000. LPBA40 resamples the five held-out subjects rather than treating its 20 ordered pairs as independent. Fixed-path comparisons reselect the constant reference within each resample. These intervals quantify cohort uncertainty for fixed checkpoints, not variation across training seeds. Update-order tests use paired Wilcoxon tests with Holm correction; scheduling intervals are unadjusted.