KNOWING WHEN NOT TO TRUST: 3D FLOW MATCHING WITH SELF-FLAGGING UNCERTAINTY FOR POST-SURGICAL CEREBRAL BLOOD FLOW PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting post-operative cerebral blood flow (CBF) from pre-operative imaging could support surgical planning for Moyamoya patients undergoing revascularization. We present two results built on that observation. First, recast as a data-to-data bridge, conditional flow matching clears the identity baseline for the first time, cutting brain-masked MAE 15.7% below it with a third of the sampling steps and beating it for 95–100% of held-out subjects on each of the four brain-masked metrics. Second, the same trained model doubles as its own trust signal: the score of its velocity field is available in closed form, so a Langevin-corrected SDE samples the model's own probability path and yields a per-voxel uncertainty map at no cost in accuracy and with no tuned noise scale. The map orders error monotonically, its most uncertain decile carrying $2.3\times$ the error of its most confident, and the ordering holds for 40 of 42 held-out subjects. Aggregated to one number per patient it is far weaker, since a subject's error mostly reflects how much that subject changed. We therefore deliver the map rather than a scalar: decision support that shows where a prediction should not be trusted, not an autonomous predictor. All experiments are reproducible and code will be released upon acceptance of paper.

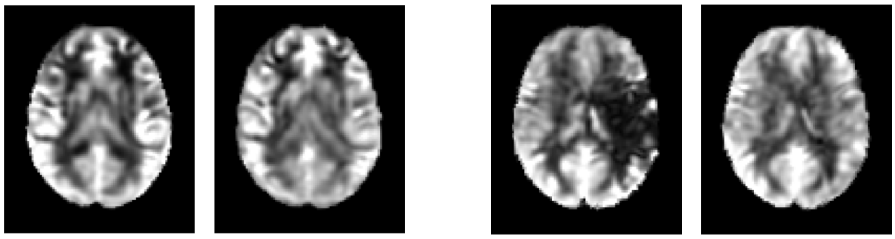
1 INTRODUCTION

Moyamoya disease progressively narrows the internal carotid arteries, reducing cerebral perfusion and raising stroke risk (Scott & Smith, 2009). Revascularization neurosurgery, the standard treatment, produces abrupt perfusion changes that are hard to anticipate. While direct or indirect bypass aims to raise cerebral blood flow (CBF) in hypoperfused territories, the postoperative response varies regionally, and focal hyperperfusion can occur (Fujimura et al., 2011). Figure 1 shows two examples of how hard this change is to predict.

A reliable predictor of post-surgical CBF could inform surgical planning and flag hemodynamic complications earlier. Unlike most longitudinal imaging tasks, the target is not gradual progression but an intervention-driven change. Existing work is largely 2D and slice-based (Zhao et al., 2025), discarding the through-plane dependencies 3D hemodynamics assume. Most such models are also opaque, offering no uncertainty signal to tell a clinician when to trust them.

We take as input a pre-surgical CBF volume x_{pre} from single-delay Arterial Spin Labeling (ASL) MRI and predict the post-surgical volume $\hat{x}_{\mathrm{post}}$. We first build a two-stage conditional 3D latent diffusion model (CLDM) and compare it against LDM, 3D U-Net, 2D MONAI, and CycleGAN baselines (Section 4.1). Although CLDM outperforms these learned baselines, it does not beat the identity predictor, which simply reuses the pre-operative scan as its prediction. This suggests the model should preserve pre-operative anatomy and perfusion by construction, learning only the intervention-induced change. We implement this with conditional flow matching (Lipman et al., 2023; Tong et al., 2024), learning a transport bridge from x_{pre} to x_{post} (Section 3.2) that cuts brain-masked MAE 15.7% below the identity function, which no diffusion variant we trained beats. This raises a clinically important question: when should a prediction not be trusted? We derive an uncertainty measure from the learned velocity field and evaluate it at voxel and subject levels (Section 3.3).

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.



x_{pre} (1) x_{post} (1) x_{pre} (2) x_{post} (2)

Figure 1: Pre- and post-surgery scans from two subjects. The left pair shows discrete changes in selected tissues; the right shows a large regional rise in perfusion. Cases like these make post-surgery CBF hard to predict by pattern matching.

2 Related work

Post-surgical CBF and longitudinal modeling. Paired CBF data and clinical motivation come from revascularization imaging work (Zhao et al., 2025); BrLP (Puglisi et al., 2025) models gradual progression via latent diffusion. Concurrent CBF work is 2D slice-based (Zhao et al., 2025) and cannot capture volumetric dependencies.

Latent diffusion. LDMs apply denoising in a compressed latent space (Rombach et al., 2022), attractive for 3D medical images where image-space diffusion is memory-prohibitive. We adopt the DDPM noise-prediction objective (Ho et al., 2020) and DDIM sampling (Song et al., 2021).

Flow matching and bridges. Flow matching regresses a velocity field onto a conditional velocity along a prescribed path (Lipman et al., 2023); Tong et al. (2024) generalize it and give the reference implementation we follow. The source distribution is a free choice and need not be noise. Rectified flow (Liu et al., 2023b) and image-to-image Schrödinger bridges (Liu et al., 2023a) transport between two *data* distributions, the structure of paired prediction and the reason a bridge sidesteps the failure in Section 4.1. Stochastic interpolants (Albergo et al., 2023) supply the result behind our ensemble: for a Gaussian path, a family of SDEs shares the probability-flow ODE's marginals.

Uncertainty for image regression. Kendall and Gal (2017) separate aleatoric from epistemic uncertainty, a distinction that explains our subject-level result. Deep ensembles (Lakshminarayanan et al., 2017) are the standard epistemic estimator but need multiple trainings. In medical imaging, repeated diffusion sampling gives per-voxel spread (Wolleb et al., 2022); our SDE sampler is the flow-matching analogue. We evaluate with sparsification and AUSE (Ilg et al., 2018). Conformal image-to-image regression (Angelopoulos et al., 2022) would upgrade correlations to guarantees.

Other baselines. The 3D U-Net (Çiçek et al., 2016) regresses voxel-wise but oversmooths; MONAI (MONAI Consortium, 2020) provides our 2D reference; CycleGAN (Zhu et al., 2017) is unstable in 3D, as are related volumetric GANs (Sun et al., 2022; Bourigault et al., 2024). The 7TCDM architecture (Wang et al., 2025) informs our 3D U-Net, held fixed across diffusion and flow models so comparisons isolate the objective.

3 Method

3.1 Conditional latent diffusion

Given paired volumes $x_{\text{pre}}, x_{\text{post}} \in \mathbb{R}^{D\times H\times W}$, we learn $p_\theta(x_{\text{post}} \mid x_{\text{pre}})$. The two-stage pipeline is standard. Stage 1 trains a KL-regularized 3D VAE with $z = \mu + \sigma \odot \epsilon$. Stage 2 freezes it and trains a conditional denoiser in latent space: $z_t = \sqrt{\bar{\alpha}_t} z_{\text{post}} + \sqrt{1-\bar{\alpha}_t}\epsilon$, the denoiser receives $[z_t, z_{\text{pre}}]$, and

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, z_{\text{pre}}, t)\|_2^2\right]. \tag{1}$$

Inference uses 50 DDIM steps from Gaussian noise with classifier-free guidance $\hat{\epsilon} = \hat{\epsilon}_\emptyset + w(\hat{\epsilon}_{\text{pre}} - \hat{\epsilon}_\emptyset)$ at $w = 3.0$; lower values collapsed toward the unconditional mean, higher ones introduced artifacts.

Every model above generates x_{post} from Gaussian noise, a framing inherited from unconditional synthesis. This formulation is poorly matched to the present task because the two scans are the same

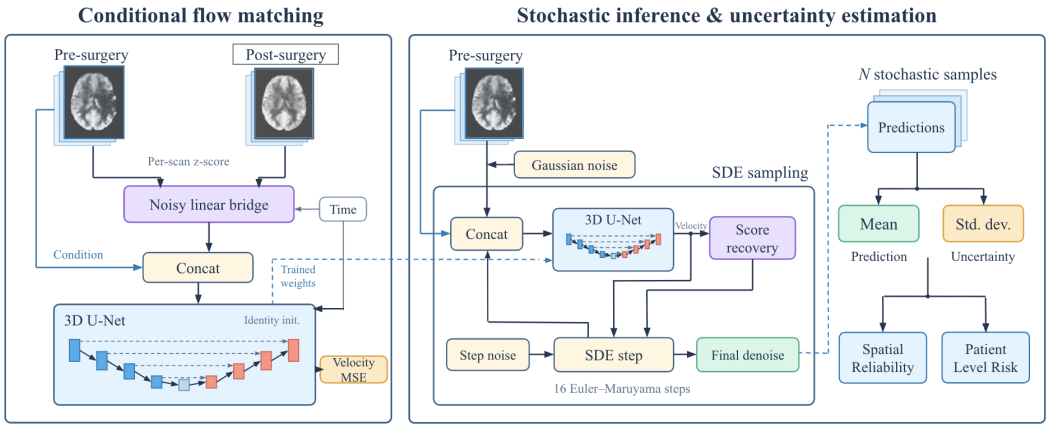


Figure 2: The two stages of the method, sharing one network. **(A, left) Conditional flow matching.** The noisy linear bridge mixes two real scans at a sampled time; the identity-initialized 3D U-Net, which also receives x_{pre} by channel concatenation, regresses the velocity $u_t = x_{\text{post}} - x_{\text{pre}}$ under an MSE loss. **(B, right) Stochastic inference and uncertainty estimation.** The same trained field is reused at inference: the Gaussian path's score is closed-form in the velocity (Eq. 4), so 16 Euler-Maruyama steps from $x_{\text{pre}} + \sigma\varepsilon$ sample the model's own probability path, and a final denoise strips the path's σ floor. Over K trajectories the mean is the prediction and the per-voxel standard deviation is the uncertainty, read as a spatial reliability map and as a patient-level risk score.

brain six months apart. Consequently, the trivial predictor $\hat{x}_{\text{post}} := x_{\text{pre}}$, which we term the *identity baseline*, is already strong, as Figure 3 shows. This is structural, not a tuning failure: round-tripping the *true* x_{post} through the frozen autoencoder scores MAE 0.149 / PSNR 31.63 / SSIM 0.965, so the latent space is not the constraint. The constraint is that generation-from-noise re-synthesizes anatomy the input already contains, which 206 paired subjects cannot fund.

3.2 Conditional flow matching as a pre-to-post bridge

Flow matching learns a velocity field transporting a source to a target along a prescribed path. The source is a free choice, which is the degree of freedom Section 4.1 says to exercise: we take it to be the pre-operative volume, making the model a *data-to-data bridge* rather than a generator (Figure 2A). Following Lipman et al. (2023); Tong et al. (2024), with $t \sim \mathcal{U}(0,1)$ and $\varepsilon \sim \mathcal{N}(0, I)$,

$$x_t = (1-t)\, x_{\text{pre}} + t\, x_{\text{post}} + \sigma\varepsilon, \qquad u_t = x_{\text{post}} - x_{\text{pre}}, \tag{2}$$

and we regress on that target, conditioning on x_{pre} by channel concatenation:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,\varepsilon}\left[\|v_\theta(x_t, t \mid x_{\text{pre}}) - u_t\|_2^2\right]. \tag{3}$$

At inference we integrate $dx/dt = v_\theta$ from $x(0) = x_{\text{pre}}$ with an 8-step Heun solver. Three choices are load-bearing; Section 4.1 measures each.

The network must see x_{pre}. The task is conditional on a specific scan, so a field given only x_t can at best learn $\mathbb{E}[u_t \mid x_t]$, which transports the marginals while scrambling the pairing. Empirically the loss then barely moves (Section 4.1).

Training starts at the identity baseline. Zero-initializing the final convolution makes $v_\theta \equiv 0$, so the untrained model returns exactly x_{pre}; training begins at the strongest simple predictor and learns only the residual. The path is short, so 8–16 function evaluations suffice against 50 DDIM steps.

Interpolant noise $\sigma > 0$ is required. At $\sigma = 0$ the path is the deterministic line between endpoints, and a network given x_t and x_{pre} reads the target off algebraically as $(x_t - x_{\text{pre}})/t$. Training loss then falls toward zero while sampling collapses: the velocity at $t = 0$ is $0/0$, so the prediction never leaves the input. Setting $\sigma > 0$ removes the degeneracy and forces the genuine conditional mean, the standard image-to-image bridge fix (Liu et al., 2023a; Albergo et al., 2023).

3.3 UNCERTAINTY AS A TRUST SIGNAL

A model that is right on average is not actionable if a clinician cannot tell which of its outputs to believe. Both answers, *where* and *when*, come from the field already trained in Section 3.2, with no ground truth or second model. Figure 2B gives the chain; two choices make it well-posed.

The ensemble samples the model's own probability path. The probability-flow ODE is deterministic, so an ensemble needs stochasticity. Perturbing $x(0)$ by $\eta\varepsilon$ supplies it, but η is an arbitrary knob trading accuracy against spread. None is needed: the training path (Eq. 2) is Gaussian, so its score follows in closed form from the learned velocity. Since $\mathbb{E}[x_{\text{post}} \mid x, t] = x_{\text{pre}} + v_\theta$,

$$s_\theta(x,t) = \nabla_x \log p_t(x \mid x_{\text{pre}}) = -\frac{x - x_{\text{pre}} - t\, v_\theta(x,t)}{\sigma^2}, \tag{4}$$

and by the stochastic-interpolants result (Albergo et al., 2023) the family $dx = [v_\theta + \gamma s_\theta]\, dt + \sqrt{2\gamma}\, dW$ shares the ODE's time marginals for every $\gamma \geq 0$. We start from the training path's own $p_0 = \mathcal{N}(x_{\text{pre}}, \sigma^2)$, take 16 Euler-Maruyama steps at $\gamma = \sigma^2$, and return the *denoised* endpoint $x_{\text{pre}} + v_\theta(x(1), 1)$ rather than $x(1)$: p_1 is the target convolved with $\mathcal{N}(0, \sigma^2)$, so the raw endpoint carries an additive σ floor that is path noise, identical everywhere and unrelated to predictive difficulty. All stochasticity is thus the trained σ. The K samples give $\bar{x} = \text{mean}_k\, x_k$ and $u = \text{std}_k\, x_k$.

Two spatial scales. With G a Gaussian of $\sigma_s = 2$ voxels,

$$\tilde{e} = |G(\bar{x} - x_{\text{post}})|, \qquad \tilde{u} = \text{std}_k\, G(x_k - x_{\text{pre}}), \tag{5}$$

alongside the raw pair $e = |\bar{x} - x_{\text{post}}|$ and u. Smoothing must act on the *signed* residual: because $G(\bar{x} - x_{\text{pre}}) - G(x_{\text{post}} - x_{\text{pre}}) = G(\bar{x} - x_{\text{post}})$, the coherent error is the blur of the signed error, which cancels sign-balanced speckle, whereas blurring $|e|$ preserves it. Symmetrically $\tilde{u}$ is the spread of the smoothed *edits*, i.e. the uncertainty of the coherent prediction rather than a smoothed copy of u. Section 4.3 reports both pairs.

4 EXPERIMENTS AND RESULTS

4.0.1 DATASET

On a machine with a double A6000 GPU, paired pre- and post-surgical 3D CBF volumes from high-risk Moyamoya patients were acquired with single-delay ASL MRI. The curated release holds $N = 235$ pairs at native $91 \times 109 \times 91$ resolution. Every scan is skull-stripped with FSL, registered to MNI152 with ANTs and z-scored, so all metrics are in z-score units on the brain foreground. Augmentation applies random left-right flips and $\pm 10^\circ$ rotations identically to both volumes of a pair. Final results use 5-fold cross-validation. Folds were drawn separately as each model family was developed, so held-out sets differ between models, and every metric is reported beside an identity baseline computed on that model's own subjects.

4.0.2 EVALUATION

Every model is scored with the same harness. The primary metrics are the conventional four (MAE, MSE, PSNR, SSIM), computed in z-score space on the brain foreground B, each reported beside the identity baseline $\hat{x}_{\text{post}} := x_{\text{pre}}$. Each scan is z-scored over the whole volume, so its background is a constant plateau rather than zero, and a nonzero test ($x \neq 0$, the convention of the latent-diffusion code we build on) would select every voxel. We instead take the voxels that differ from the modal value in either scan, $B = \{v : x_{\text{pre}}(v) \neq \text{mode}(x_{\text{pre}})\} \cup \{v : x_{\text{post}}(v) \neq \text{mode}(x_{\text{post}})\}$, which covers about 29% of the volume. We also record the per-subject win rate over that baseline, which Figure 3 plots. Flow models are scored with the deterministic 8-step Heun ODE; the stochastic sampler of Section 3.3 is used only for the uncertainty analysis (Section 4.3).

Means are a blunt instrument here even inside B: most of the brain is parenchyma the surgery left unchanged, where every method ties, so a large gain on the surgical edit barely moves the mean. Over the whole volume, roughly 71% is background. We therefore add a third family of metrics restricted to the change ROI R, the top f=5% of B by $|\Delta|$ where $\Delta = x_{\text{post}} - x_{\text{pre}}$, and report it in two forms.

Raw and coherent change. A sizeable part of Δ is not surgery. The two volumes are separately acquired scans warped to a template, so sub-voxel misalignment along any sharp boundary produces large $|\Delta|$ nothing in x_{pre} predicts. Two diagnostics size it: R is $\approx 2.1\times$ enriched on the top-20% pre-operative gradient-magnitude voxels, and with $\Delta_{\mathrm{coh}} = G_\sigma[\Delta]$, $\Delta_{\mathrm{hf}} = \Delta - \Delta_{\mathrm{coh}}$ at σ=2 voxels,

$$\texttt{coherent_frac} = \frac{\|\Delta_{\mathrm{coh}}\|_R^2}{\|\Delta_{\mathrm{coh}}\|_R^2 + \|\Delta_{\mathrm{hf}}\|_R^2} \approx 0.43,$$

so over half the raw ROI energy sits in Δ_{hf}. The harness reports `change_mae_improvement`, scored against Δ, and `coherent_mae_improvement`, scored against Δ_{coh} with the prediction smoothed to the same scale. Identity scores exactly 0 on both by construction.

No block substitutes for the others. The raw ROI metric is the least forgiving and partly measures noise; the coherent one credits only the recoverable low-frequency edit but ignores fine detail; the four conventional metrics are what every prior method reports. Figure 3 shows all three, read together: a model that wins one block and loses another is informative, and Section 4.1 has such a case.

4.1 Model comparison

Figure 3 scores fifteen checkpoints in three blocks. Each dot is a held-out subject, the open circle its mean. The top block is absolute score (MAE and MSE lower is better, PSNR and SSIM higher), with a grey diamond at that model's identity baseline. The middle and bottom blocks are improvement over that baseline, on the whole brain and on the coherent change inside R; there higher is always better, zero is identity, and the annotation gives the share of subjects that beat it. Only the improvement blocks compare models fairly, because the diamonds do not line up: each model was scored on its own held-out subjects, so absolute error confounds model quality with split difficulty. A common test set would fix this; we do not have one due to shortage of data, which is addressed in limitations.

The conditioned bridges win, and the source distribution is why. `flow-cfm` beats identity for $100/100/100/95\%$ of subjects across the four metrics and `flow-sharp` for $100/100/100/97\%$; the selected bridge reaches brain-masked MAE 0.405 against an identity baseline of 0.480 (-15.7%). Scored over the whole volume instead, which adds each scan's background plateau to the mean, the same model reaches 0.213 against 0.250 and beats identity on MAE for 71% of subjects. We report the brain-masked figures because the background carries no perfusion signal. The controlled comparison is `flow-noise-cfm`, identical but for starting from Gaussian noise: $7/3/3/10\%$. Generation from noise re-synthesises anatomy the input already contains, which 206 pairs cannot fund. Conditioning is what makes the objective descend at all: without x_{pre} on the input the training loss moves $0.150 \to 0.147$ over eight epochs, against $0.121 \to 0.037$ with it.

Capacity does not rescue the wrong framing. `cldm-base`, `ldm3d`, `cdm3d` and `cyclegan3d` look respectable in the top block and beat identity for 0% of subjects in the middle; a table without the identity row would rank them as working methods. Scaling CLDM (larger denoisers, richer latents, stronger guidance) reaches at best $34/51\%$ on MAE/MSE, still below every bridge.

The blocks disagree about one model, by design. `flow-nd-bridge` trains with a change-weighted MSE, sits at 70% on whole-brain MAE, and posts the largest coherent-change improvement in the bottom block at 100% of subjects. Up-weighting the ROI buys exactly what it optimises and costs whole-brain error. `cyclegan3d`, at 0% in every block, is the check that the bottom block is not simply generous to everything.

4.2 Qualitative behaviour

Figure 4 shows a flow-matching prediction on a held-out subject, with three zoomed regions.

- **Important surgery-driven change signals are learned.** In regions A and B the pre-operative scan is comparatively flat, the ground truth develops clear territory-scale perfusion structure, and the prediction reproduces that structure at the same location and extent. The bridge starts at x_{pre}, so a model that learned nothing would return the first column; these panels are the visual form of the win rates in Figure 3.
- **Details are given attention.** Region C holds a bright posterior increase, present in the target and absent pre-operatively, and the prediction places it correctly. The predicted change map

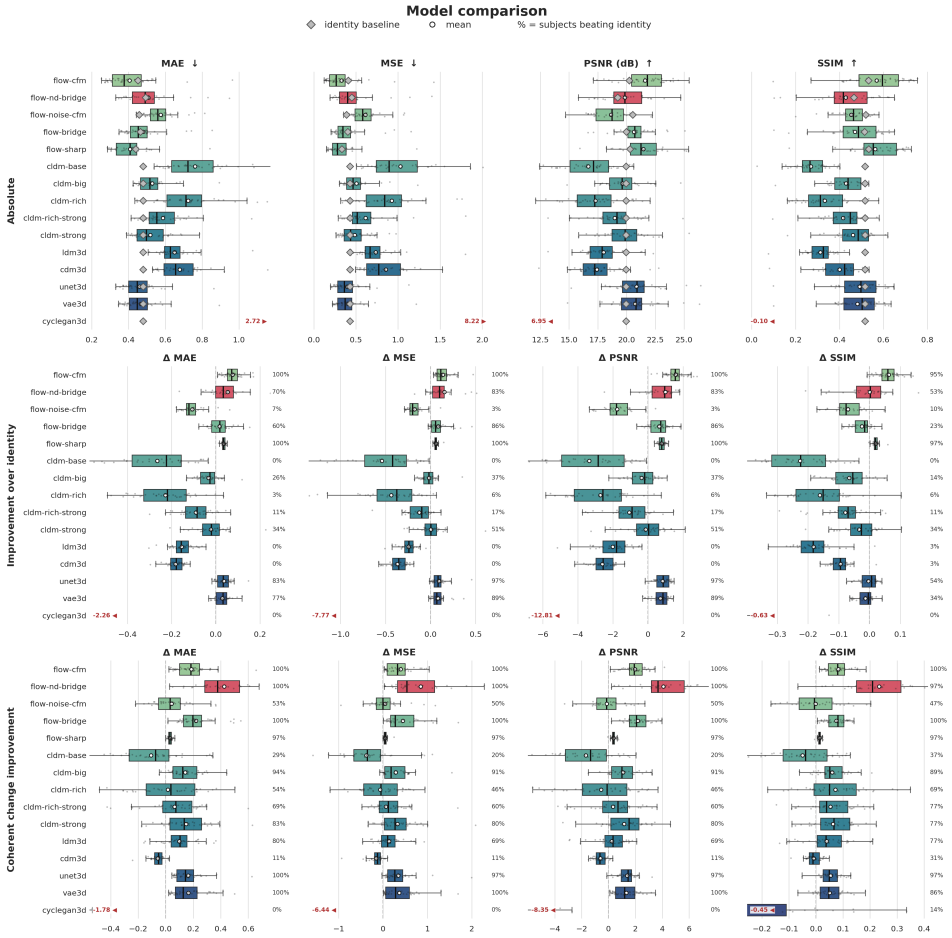


Figure 3: Model comparison; each dot is a subject and the open circle the mean. All metrics are brain-masked, and flow models use the deterministic 8-step Heun sampler. **Top:** absolute scores, with the grey diamond marking each model's own identity baseline. The diamonds differ because models were scored on different splits. **Middle:** improvement over that baseline, annotated with the share of subjects for which the model beats copying the pre-operative scan. **Bottom:** the same improvement measured on the coherent change inside the top-5% change ROI, which removes the misregistration component. Models that look competitive in the top block but sit at 0% in the middle one are beaten by doing nothing.

correlates with the true change at $\rho = 0.53$ on average and above 0.3 for 27 of 30 held-out subjects, with a predicted-to-true amplitude ratio of 0.96. A copy-paste predictor would sit near 0, a timid one well below 1.

- **The characteristic residual failure is contrast, not location.** Across the set the prediction is spatially right and lower in contrast than the ground truth, most visible in region C where the predicted increase is smoother and broader than the true one. This is the expected behaviour of a conditional mean under a squared-error objective, and it is why SSIM separates models MAE ranks together.

4.3 Uncertainty analysis

We draw $K{=}20$ trajectories per subject with the sampler of Section 3.3 and bin brain voxels into deciles of predicted uncertainty (Figure 5). The sampler choice is settled first: perturbing the deterministic ODE's start by $\eta\varepsilon$ trades directly against accuracy (MAE $0.1835 \to 0.2098$ as η goes $0.05 \to 0.3$) and gives the weaker signal at matched cost, a raw sparsification gain of 19.7% against 37.2%. All results in this section use the score-derived SDE, which has no such knob. On `flow-cfm`

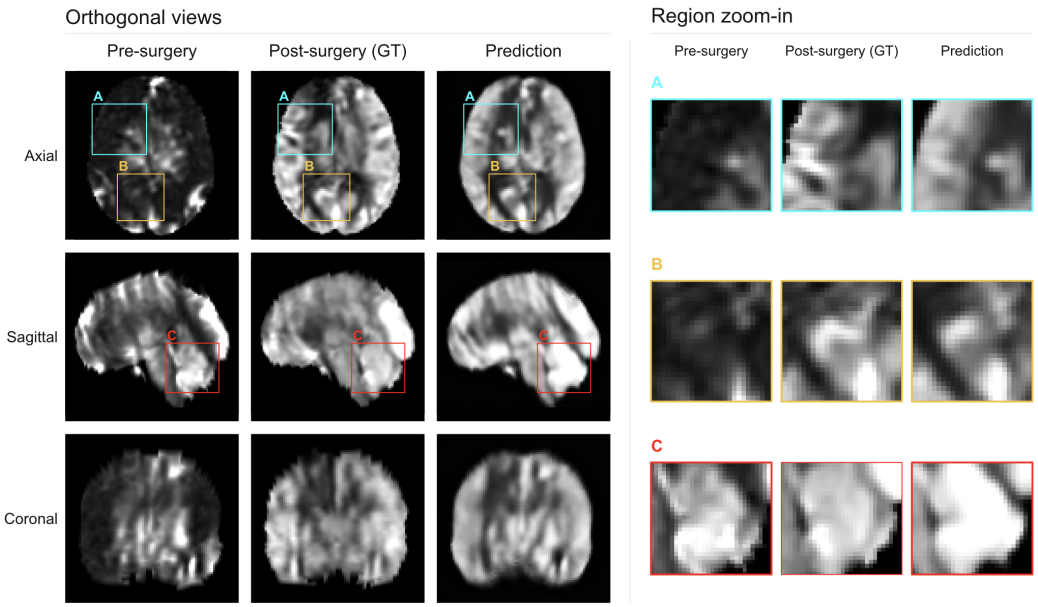


Figure 4: Qualitative prediction for a held-out subject. **Left:** pre-operative input, post-operative ground truth and prediction in three orthogonal views. **Right:** zoomed regions A to C. The model introduces perfusion structure that is absent from the input and present in the target, at the correct location and extent, but at reduced contrast.

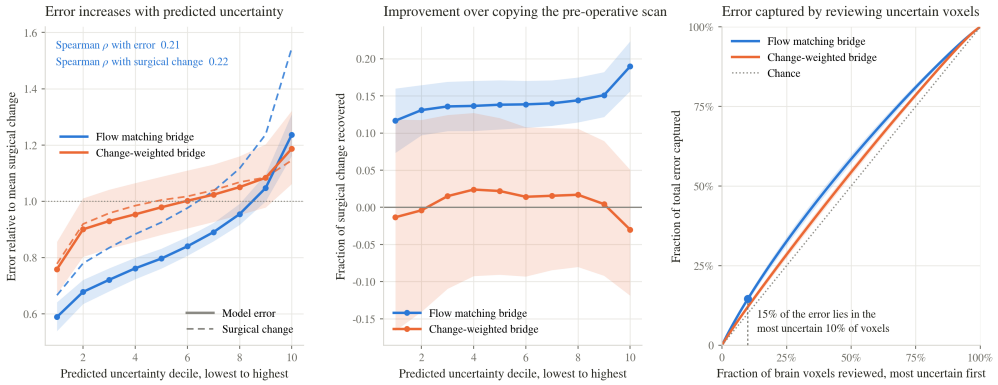


Figure 5: Uncertainty behaviour for the two bridge models; bands are 95% bootstrap intervals over subjects. **Left:** binning voxels by predicted uncertainty orders both the model's error (solid) and the actual surgical change (dashed). **Centre:** the gap between them as a fraction of the local surgical change; positive means better than copying x_{pre}, and the change-weighted model turns negative in its least confident decile. **Right:** share of total error captured against the fraction of voxels reviewed, most uncertain first. Curves are normalised per subject by that subject's mean surgical change, which makes the two models comparable despite different folds.

its ensemble mean reaches brain-masked MAE 0.401 against identity's 0.480 and beats identity for all 42 subjects, so it costs no accuracy relative to the ODE of Figure 3.

Error rises with predicted uncertainty (left). Mean error rises monotonically across all ten deciles for both bridges. At the coherent scale the most uncertain decile carries $2.27\times$ the error of the most confident for `flow-cfm` and $1.64\times$ for `flow-nd-bridge`, and per-subject ρ is positive for 40 of 42 and 29 of 30 subjects (sign test $p \approx 10^{-10}$). The dashed curves rise too, revealing what the map has learned: uncertainty tracks the true surgical change ($\rho = 0.22$) as strongly as it tracks the model's own error ($\rho = 0.21$). It marks where surgery acted, and the model is worst where surgery acted.

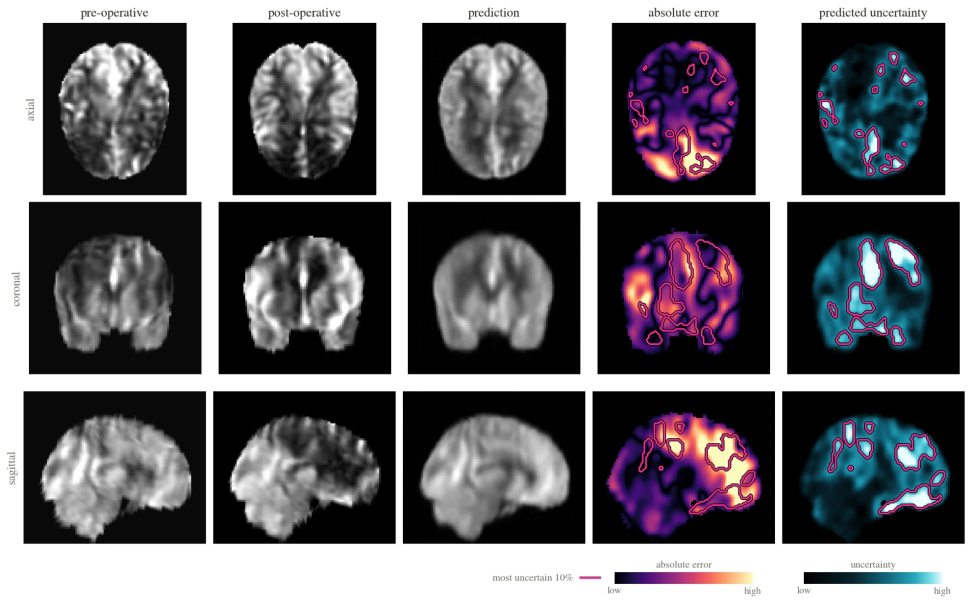


Figure 6: One held-out subject in three orthogonal views: the pre-operative scan, the post-operative ground truth, the prediction, the resulting error, and the model's predicted uncertainty. The prediction flattens much of the real pre-to-post change, so the error is large. The magenta outline is the most uncertain tenth of the brain, thresholded from the ensemble spread alone, so available before ground truth exists; it holds 35% of the largest errors against 10% expected by chance. Uncertainty uses a separate colour scale. The claim is spatial co-localisation, not equal magnitude.

The gap between the curves is the model's value (centre). The coherent surgical change is exactly the coherent error of copying x_{pre}, so the dashed curve is both "where surgery acted" and "the bar to beat". As a fraction of the local change, `flow-cfm` rises from 0.12 to 0.19: it contributes most where it is least sure. `flow-nd-bridge` is flat near zero and turns negative in its least confident decile, worse than copying the input. A whole-brain mean ranks them close; this curve separates them.

The operational lift is modest (right). Reviewing the most uncertain 10% of voxels captures 15% of the total error, a $1.5\times$ lift over chance. We plot the coverage curve rather than quote the decile ratio, which flatters the method by contrasting two extremes.

Two deflations belong with these numbers. Coherent risk ratios sit below raw ones (2.27 against 2.82; 1.64 against 2.02; paired, $p < 10^{-8}$), so part of any naive uncertainty-error correlation is the ensemble anticipating registration speckle no model can predict away. And the per-subject aggregate does not survive its confound: subject error is almost entirely explained by how much that subject changed ($\rho = 0.94$ against the identity error), leaving a partial correlation of 0.08 ($p = 0.63$). The subject-level score is a change-magnitude detector, as the construction implies: one weight set gives aleatoric spread and no epistemic component (Kendall & Gal, 2017; Lakshminarayanan et al., 2017).

5 From uncertainty to clinical decision support

Consider what a surgeon is actually handed. Figure 6 shows one held-out patient: the scan taken before surgery, the scan taken six months after, and the model's prediction of that second scan. The prediction is wrong here in a way that matters. Between the two real scans, blood flow changes substantially across the back and right of the brain; the prediction keeps the shape but flattens most of that change, and the error map in the fourth column shows it: large, bright, and sitting in exactly the territory a surgeon would ask about.

Nothing in the first three columns reveals this. A clinician has no later scan to compare against; that is the point of a prediction. Shown on its own, the third column invites the same trust everywhere, and this is the patient for whom trusting it everywhere would be wrong.

The fifth column changes that. It is computed from the model alone, before any post-operative scan exists, by asking it the same question twenty times and measuring how much the answers disagree. The magenta outline marks the tenth of the brain where they disagree most. It is drawn with no access to the error map beside it, and it lands on the error: it holds 35% of this patient's largest errors against the 10% chance would give, with a mean error three times that of the most confident tenth. The model does not know the right answer here, but it knows that it does not know.

Show the uncertainty with the prediction. Render the map so that it visibly de-emphasises what the model cannot support. This changes the question from "is this prediction correct?", which a clinician cannot answer, to "is this a region the model handles well?", which the map does answer. The stronger direction is confidence rather than alarm: the most confident tenth carries roughly half the typical error, so the defensible claim is about where the map *can* be relied on.

Order the reading queue. For predictions computed ahead of a planning conference, uncertainty ranks the worklist: at a 10% review budget a reader sees 15% of the total error. That is attention allocation, not a safety guarantee, and the coverage curve keeps the trade-off visible at any budget.

Choose the model on this curve. The two bridges fail in opposite directions (Figure 5, centre): the change-weighted model becomes worse than doing nothing exactly where it is least confident, while the plain bridge helps most there. A whole-brain error average ranks them close together and would miss this. A model that degrades where it is unsure is the one that should not reach a patient.

What we do not offer is a single confidence number per patient. Collapsing the map to one score mostly measures how much that patient changed rather than how wrong the model is (Section 4.3), and a system displaying such a number would assert a reliability our measurements do not support. The map is the robust product, and it should be delivered as a map.

6 Discussion and limitations

The framework was never the binding constraint; the framing was. Generation from noise loses to copying the input (Figure 3, middle block), and the fix was not a better denoiser but a different source distribution. The lesson should transfer to other paired medical prediction tasks where input and target are the same anatomy at two timepoints: the trivial predictor is strong, must be reported, and a generative model should start from it.

Limitations. *Sample size:* High-quality perfusion scans are hard to obtain, so data are scarce. Held-out folds of 30–42 subjects support the voxel-level claims, which aggregate hundreds of thousands of voxels per subject, far better than per-patient claims, whose intervals are wide. *No epistemic uncertainty:* one shared set of weights captures aleatoric spread along the sampling path and no model uncertainty (Kendall & Gal, 2017); per-patient failure is largely epistemic, and deep ensembles (Lakshminarayanan et al., 2017) are the missing ingredient. *Ground-truth quality:* some pairs differ in coverage or registration and dominate the error tail; where the target is unreliable those cases are arguably mislabelled rather than mispredicted. *Correlational, not calibrated:* we validate by rank correlation, not coverage; conformal image-to-image regression (Angelopoulos et al., 2022) would give intervals, though loose given our calibration set. *Imaging-only conditioning:* the model ignores clinical covariates (age, severity, surgical approach, hemisphere) available in this cohort.

7 Conclusion

Every generation-from-noise model we trained is beaten by copying the pre-operative scan. Recasting the task as a conditional flow matching bridge from x_{pre} produces the first models that clear that baseline, for 100% of held-out subjects on three of four brain-masked metrics. The same field yields, at no cost in accuracy, a per-voxel uncertainty map that orders error monotonically and flags regions with up to $3\times$ the error of confident ones; reviewing a tenth of the voxels captures 15% of the error. The result is decision support with an explicit competence map, not an autonomous predictor.

AI USE STATEMENT

AI is used extensively for coding and experiment design. However, most of the writeup are directed by the authors, with AI checking grammar and typos, giving suggestions for specific wording.

REPRODUCIBILITY STATEMENT

All experiments are reproducible with code in Github that is released upon acceptance.

REFERENCES

M. S. Albergo, N. M. Boffi, and E. Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. In *arXiv preprint arXiv:2303.08797*, 2023.

A. N. Angelopoulos, A. P. Kohli, S. Bates, M. I. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, and Y. Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, 2022.

E. Bourigault, A. Hammers, and A. Wachinger. X-diffusion: Generating detailed 3D MRI volumes from a single image using cross-sectional diffusion models. *arXiv preprint arXiv:2405.09746*, 2024.

Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention*, 2016.

M. Fujimura, H. Shimizu, T. Inoue, S. Mugikura, A. Saito, and T. Tominaga. Significance of focal cerebral hyperperfusion as a cause of transient neurologic deterioration after EC-IC bypass for moyamoya disease. *Neurosurgery*, 68(4):957–964, 2011.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

E. Ilg, Ö. Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox. Uncertainty estimates and multi-hypotheses networks for optical flow. In *European Conference on Computer Vision*, 2018.

A. Kendall and Y. Gal. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 2017.

B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

G.-H. Liu, A. Vahdat, D.-A. Huang, E. A. Theodorou, W. Nie, and A. Anandkumar. I^2SB: Image-to-image Schrödinger bridge. In *International Conference on Machine Learning*, 2023a.

X. Liu, C. Gong, and Q. Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023b.

MONAI Consortium. MONAI: Medical open network for AI, 2020. `https://github.com/Project-MONAI/MONAI`.

L. Puglisi et al. Medical image prediction for longitudinal and intervention-driven imaging. *Medical Image Analysis*, 2025.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

R. C. Scott and P. J. Smith. Moyamoya disease and moyamoya syndrome. *New England Journal of Medicine*, 360(12):1226–1237, 2009.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Q. Sun et al. HA-GAN: Hierarchical amortized training for 3D brain MRI generation. *Frontiers in Neuroscience*, 16, 2022.

A. Tong, K. Fatras, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Y. Wang et al. Conditional diffusion modeling for high-resolution neuroimaging. *bioRxiv*, 2025.

J. Wolleb, R. Sandkühler, F. Bieder, P. Valmaggia, and P. C. Cattin. Diffusion models for implicit image segmentation ensembles. In *Medical Imaging with Deep Learning*, 2022.

M. Zhao et al. Paired cerebral blood flow imaging for revascularization neurosurgery. *Journal of Cerebral Blood Flow & Metabolism*, 2025.

J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.