This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Conditional Flow Matching for Local Synthesis of Healthy Brain Tissue in 3D MRI

Bui Van Hung[1], Dang Thanh Vu[2], Hoang-Son Vo[1], Mai Duc Minh Huy[3], Mai Tien Dung[4], Quoc-Tien Nguyen[1], Hyung-Jeong Yang[1], and Soo-Hyung Kim[1]

[1] Chonnam National University, Gwangju, South Korea
{hungbv16, hoangsonvothanh, tiennq27, hjyang, shkim}@jnu.ac.kr
[2] AISEED, Inc
dtvu@aiseed.kr
[3] Ho Chi Minh University of Science, Ho Chi Minh City, Vietnam
minhhuymaiduc@gmail.com
[4] Institute for Cybersecurity and Digital Technologies, Moscow, Russia
may.t@edu.mirea.ru

Abstract. Many brain-MRI tools assume a healthy, intact brain, so they fail on scans that contain tumours or surgical cavities. The BraTS 2026 Local Inpainting Challenge targets this gap: given a scan with a masked-out region, the task is to synthesise realistic healthy tissue that agrees with the surrounding anatomy. We address it with a conditional flow-matching model that learns a near-straight transport from noise to brain anatomy from a single, stable training loss. To keep every fill patient-specific, the surrounding scan conditions the network both as additional input channels and through a spatially adaptive modulation pathway that steers features at every scale. We also align the training loss with the scoring protocol by concentrating supervision on the evaluated healthy region while keeping a valid global velocity field, and enlarge the training data with the challenge's own healthy-mask placement procedure. A central observation drives the inference stage: the squared-error-based metrics reward the average of plausible fills rather than perceptual sharpness. We exploit this with a variance-reduction recipe combining test-time flip augmentation with multi-sample averaging, which markedly improves fidelity on the hardest large-mask cases without any retraining. Measured with SSIM, PSNR, and MSE inside the healthy region, our method produces coherent, natural-looking tissue, surpasses learned image-to-image baselines, and both trains and runs on a single commodity GPU.

Keywords: Brain MRI · Inpainting · Flow matching · Generative models · Image synthesis · BraTS.

1 Introduction

Many automatic brain-MRI pipelines, from registration to segmentation and atlas-based analysis, assume a healthy, intact brain, an assumption broken by

tumours and by the surgical cavities that follow treatment. A practical fix is to inpaint the affected area with realistic healthy tissue so that standard tools still apply [12]. The BraTS 2026 Local Inpainting Challenge [12] frames this as synthesis: each case gives a T1-weighted scan with a region masked out and its binary mask, and the goal is to fill the mask with intensities that match the surrounding anatomy. Only a separate patch of healthy tissue inside the mask is scored, using SSIM [21], PSNR, and MSE, because a healthy reference exists only there. Data come from the BraTS-GLI 2023 cohort [2, 16, 5, 3, 4]. Because the scored patch sits away from the disease, the protocol grades tissue well inside the hole, where a fill obtained by smoothly continuing the mask boundary is unlikely to suffice.

We adapt conditional flow matching to this task. Diffusion models [8, 20, 19, 15] give strong image quality, but their long sampling chains are costly in 3D. Flow matching [13, 14, 1] instead trains a network to predict the velocity of a probability-flow ODE from one stable loss, a good fit for 3D MRI on a limited budget. The velocity field is a 3D U-Net [18, 6], and our main design choice is how the surrounding scan conditions it. Input concatenation alone is weak, as its influence fades in the deeper layers that set the fill's global structure; we therefore also route the scan through a spatially adaptive pathway inspired by SPADE [17] that modulates features at every scale. We further align the objective with the scoring rule by up-weighting the evaluated region while keeping a valid global velocity field, and diversify supervision by regenerating healthy targets with the challenge's own placement procedure [12].

A complementary insight concerns scoring. MSE is provably minimised by the conditional mean of the plausible fills, and PSNR with it, so steering a sample toward perceptual sharpness moves it off that mean and *lowers* the score; SSIM carries no such guarantee, but we observe it to improve in step. The tension is sharpest on large masks, where a single stochastic fill varies widely. We turn it into an advantage: at inference, and without retraining, we average fills over test-time flips and several noise samples, recovering the conditional mean the metrics reward.

Contributions. (i) A conditional flow-matching method for local inpainting of healthy tissue in 3D brain MRI, with a single stable loss, conditioned on the surrounding scan both by input concatenation and by SPADE-style spatially adaptive modulation at every scale. (ii) A scoring-aligned weighted objective that concentrates supervision on the evaluated region while preserving a valid global velocity field, plus a mask-regeneration augmentation reusing the challenge's own placement procedure. (iii) An analysis of the perception-distortion tension, turned into a retraining-free variance-reduction recipe (flip augmentation with multi-sample averaging) that gains most on the hardest large-mask cases and surpasses learned image-to-image baselines on a single commodity GPU. Code, configurations, and the inference recipe reproducing our submission are publicly available at `https://github.com/Hung-3008/brats_2026_simflow`.

2 Methods

2.1 Problem setup

Let x denote a T1-weighted volume, $m \in \{0,1\}$ the full inpainting mask, and $h \subseteq m$ the healthy sub-region on which the submission is scored. The input is the voided scan $\tilde{x} = x \odot (1-m)$ together with m, and the model must predict the intensities inside m. A healthy ground truth exists only under h, so we supervise and evaluate on h while still driving the model with the full-mask conditioning it sees at test time. We work on 96^3 patches centred on h; any tumour cavity inside is voided as well, matching the inference distribution. Intensities are scaled by the volume maximum and mapped to $[-1,1]$, and masks remain binary.

2.2 Conditional flow matching

We model inpainting as a conditional probability-flow ODE that transports Gaussian noise to healthy anatomy [13, 14, 1]. For a clean patch x_1 and noise $x_0 \sim \mathcal{N}(0, I)$ we use the optimal-transport interpolant

$$x_t = \big(1 - (1-\sigma_{\min})\, t\big)\, x_0 + t\, x_1, \qquad u_t = x_1 - (1-\sigma_{\min})\, x_0, \tag{1}$$

with $t \in [0,1]$ and $\sigma_{\min} = 10^{-4}$, and train a network $v_\theta(x_t, t, c)$ to regress the target velocity u_t. The conditioning $c = [\tilde{x}, m]$ is concatenated onto x_t, so a single velocity channel is predicted from three input channels, and at inference we integrate $\mathrm{d}x/\mathrm{d}t = v_\theta$ from x_0 with an explicit Euler scheme. Following RePaint-style compositing [15, 19], we keep the visible tissue verbatim, $\hat{x} = \tilde{x} \odot (1-m) + x_1^{\text{gen}} \odot m$, so the model is trusted only inside the hole.

2.3 Spatially-adaptive conditioning

The velocity field is a 3D U-Net [18, 6] with four resolution levels (channel multipliers $1, 2, 2, 4$; base width 32 for the initial model and 48 for the final system), group normalisation, and self-attention at the coarsest level only, where it is affordable in 3D. Supplying the condition by input concatenation alone is weak guidance: after the first convolutions the voided scan and mask are entangled with the noisy state, and their signal is washed out by the successive normalisations of the deeper layers, exactly those that decide the fill's global structure.

We therefore add a spatially adaptive conditioning pathway in the spirit of SPADE [17], the central architectural component of our model (Fig. 1). A lightweight condition encoder maps the raw condition $[\tilde{x}, m]$ to a multi-resolution feature pyramid aligned with the U-Net's spatial grids. Inside every residual block, immediately after group normalisation, the activations f are re-modulated by a per-voxel scale and shift,

$$\mathrm{SPADE}(f \mid c) \;=\; \gamma(c) \odot \frac{f - \mu(f)}{\sigma(f)} \;+\; \beta(c), \tag{2}$$

where the maps $\gamma(c)$ and $\beta(c)$ are predicted convolutionally from the pyramid level matching the resolution of f. Because γ and β are spatial fields rather than global vectors, the surrounding tissue and the mask geometry steer the generation *location by location* at every depth, so patient-specific context can never be normalised away. The network stays compact (15M parameters at width 32, 30M at width 48), keeping training and inference within a single 24 GB GPU.

2.4 Training objective

The challenge scores only h, yet a flow-matching model must learn a valid velocity field everywhere for the ODE to integrate, so masking the loss to zero outside h would break sampling. We instead use a spatially weighted velocity loss, normalised by the total weight so that its scale stays comparable to the plain (unweighted) flow-matching loss,

$$\mathcal{L}_v = \mathbb{E}_{t,x_0,x_1}\left[\frac{\sum_i w_i \left(v_\theta(x_t,t,c)_i - u_{t,i}\right)^2}{\sum_i w_i}\right], \qquad w = (1-m) + \lambda\, h, \tag{3}$$

where i indexes voxels. Since $h \subseteq m$, the three regions receive $w = 1$ on visible tissue outside the mask (ordinary supervision, learn to copy), $w = \lambda$ on the scored healthy region (with $\lambda = 5$), and $w = 0$ inside the unscored tumour cavity, which carries no healthy ground truth. The model therefore keeps a coherent global field while spending its capacity where it is graded.

Because velocity regression alone is indifferent to structure, the final system adds an auxiliary reconstruction term on the clean-image estimate. For the optimal-transport interpolant the endpoint is recoverable in closed form from a single network evaluation, without integrating the ODE:

$$\hat{x}_1 \;=\; a\, x_t + (1 - a\, t)\, v_\theta(x_t, t, c), \qquad a = 1 - \sigma_{\min}. \tag{4}$$

Writing $\bar{x} = (\hat{x}_1 + 1)/2$ and $\bar{y} = (x_1 + 1)/2$ for the two volumes rescaled to $[0,1]$, and SSIM_h for the SSIM map averaged over the scored voxels h (the scored metric itself), the auxiliary term is

$$\mathcal{L}_{\mathrm{aux}} = \lambda_1\, \mathbb{E}\left[\frac{\sum_i t\, w_i\, |\bar{x}_i - \bar{y}_i|}{\sum_i t\, w_i}\right] \;+\; \lambda_s\, \frac{\mathbb{E}\big[\, t\,(1 - \mathrm{SSIM}_h(\bar{x},\bar{y}))\big]}{\mathbb{E}[t]}, \tag{5}$$

and the total objective is $\mathcal{L} = \mathcal{L}_v + \mathcal{L}_{\mathrm{aux}}$ with $\lambda_1 = \lambda_s = 1$. Both terms reuse the weighted-mean normalisation of Eq. (3) and are scaled linearly by t, so the model is only asked to match the clean image when its prediction is near-clean ($t \to 1$), not when it reconstructs from near-pure noise. This supervises the quantity actually scored rather than the velocity producing it.

2.5 Data, augmentation, and optimisation

Training uses the 1,251 BraTS-GLI 2023 cases [2, 16, 5, 3, 4] released for the challenge, with no other data and no pre-trained weights. To enlarge and diversify

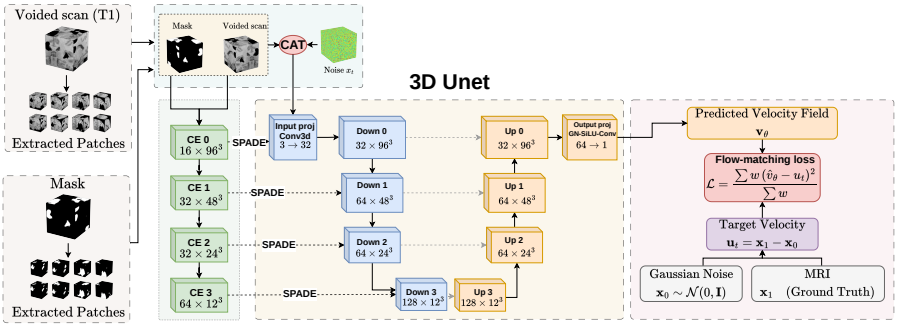

Fig. 1. Training. A clean healthy patch x_1 is interpolated with Gaussian noise x_0 to form x_t, concatenated with the condition $c = [\tilde{x}, m]$ and passed to the SPADE-conditioned 3D U-Net v_θ; a lightweight condition encoder produces a feature pyramid that re-modulates every residual block. The network regresses the optimal-transport velocity u_t under the spatially weighted loss $w = (1-m)+\lambda h$, which supervises visible tissue, up-weights the scored healthy region, and ignores the tumour cavity.

the supervision, we regenerate additional healthy masks with the challenge's own placement procedure [12], sampling further tumour-avoiding regions per subject, and apply random axis flips. We optimise with AdamW at a peak learning rate of 2×10^{-4} under a linear warm-up and decay schedule, keep an exponential moving average (decay 0.9999) of the weights, and train for 100 epochs. The final system trains the width-48 network for 100 epochs and refines it for 30 more, combining the auxiliary reconstruction term with a fine-tuning stage that adapts the sampler to a small, fixed number of function evaluations.

2.6 Inference-time variance reduction

At test time the healthy sub-region is unknown, so we blindly fill the whole mask (Fig. 2). We tile the mask bounding box with overlapping 96^3 patches, the patch size seen in training, sample each with the conditional ODE, and average the predictions across tiles before compositing onto the visible scan and rescaling to the original intensity range. The final system uses stride 32, so every voxel receives up to 27 votes, and weights each tile by an isotropic Gaussian ($\sigma = 0.25$ of the patch size), suppressing seams.

Overlap-averaging already exploits the defining property of the task: the expected squared error $\mathbb{E}\big[(\hat{x} - x)^2 \mid c\big]$ is minimised by $\hat{x} = \mathbb{E}[x \mid c]$, so averaging independent stochastic fills is optimal for MSE and PSNR, whereas sharpening an individual fill moves it off that mean and lowers those scores. SSIM is not a squared-error metric and no such optimality argument applies to it, but we empirically observe it to rise together with PSNR under every averaging source below (Table 3). We push this further with two *orthogonal* sources of averaging. *Test-time flip augmentation* (TTA) reflects the input across the volume

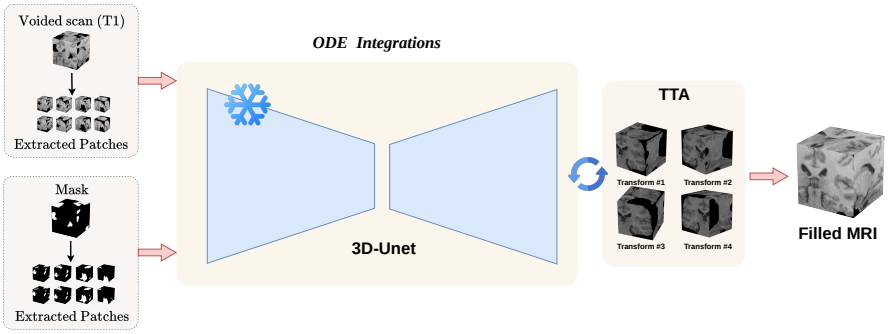

Fig. 2. Inference. The mask bounding box is tiled with overlapping 96^3 patches (stride 32), each inpainted by integrating the conditional ODE $\mathrm{d}x/\mathrm{d}t = v_\theta$. Predictions are combined through three sources of averaging, all targeting the conditional mean of plausible fills: Gaussian-weighted averaging across tiles, flip augmentation, and multi-sample ensembling over noise seeds. The averaged fill is composited onto the visible tissue and rescaled to the original intensity range.

axes, inpaints each orientation, and averages the un-flipped results, cancelling the network's orientation-dependent bias; *multi-sample ensembling* draws several independent noise seeds per orientation, cancelling the stochastic variance of the ODE. Acting on different variance sources, their gains compound. Our deployed configuration uses 4 function evaluations per patch, made possible by the few-step fine-tuning above, and combines Gaussian-blended 96^3 tiling at stride 32, eight-way flip TTA, and a two-sample ensemble ($2 \times 8 = 16$ passes).

3 Results

3.1 Experimental setup

We train on the 1,251 BraTS-GLI 2023 T1-weighted volumes and evaluate on the official BraTS 2026 Local-Inpainting validation set through the challenge's evaluation server, which scores only the healthy sub-mask. We report the three challenge metrics, SSIM [21], PSNR, and MSE, as mean $\pm$ std over the 219 validation cases under the challenge's intensity normalisation. All models train and run on a single 24 GB GPU (RTX 3090). The initial model integrates the ODE in 25 Euler steps; the final system, fine-tuned for few-step sampling, uses 4 function evaluations per patch. Training it costs about 20 GPU-hours; one tiled pass takes $\approx$5 s per case and the deployed 16-pass recipe $\approx$80 s.

Baselines. The VAE [11] and 3D Pix2Pix [10] rows of Table 1 are the reference implementations distributed by the challenge organisers.[5] We neither retrained

[5] `https://github.com/BraTS-inpainting/2026_challenge`

Table 1. Main results on the official BraTS 2026 Local-Inpainting validation set (healthy sub-mask; mean ± std). Our final system combines the wider backbone, few-step (4-evaluation) sampling, Gaussian-blended 96^3 tiling at stride 32, eight-way flip augmentation and a two-sample ensemble. For reference on a different evaluation set, the 3D Pix2Pix entry of the BraTS 2023 challenge reached ≈ 0.77 SSIM / 16.3 dB PSNR [12].

Method	SSIM ↑	PSNR ↑	MSE ↓
VAE [11]	0.750 ± 0.134	19.10 ± 3.24	0.015 ± 0.010
3D Pix2Pix [10]	0.755 ± 0.128	19.36 ± 2.96	0.014 ± 0.008
Ours, single-sample baseline	0.741 ± 0.106	19.87 ± 3.11	0.013 ± 0.008
Ours, + variance-reduced sampling	0.810 ± 0.084	21.55 ± 3.43	0.009 ± 0.006
Ours, final system	$\mathbf{0.862 \pm 0.090}$	$\mathbf{24.87 \pm 4.55}$	$\mathbf{0.0047 \pm 0.0035}$

nor re-tuned them: we ran the released code *and* checkpoints unchanged over the same 219 cases and scored them through the same official evaluation server as our own submissions. The comparison is therefore not matched on inference budget: the baselines run a single deterministic forward pass, with *no* tiling ensemble, flip TTA or multi-sample averaging, whereas our final row uses 16. Our single-sample row is the budget-matched comparison; the recipe of Sec. 2.6 is architecture-agnostic and would likely lift the baselines too, so we do not attribute the full margin to flow matching alone.

3.2 Main results

Table 1 reports our results on the official validation set, where two effects compound. First, inference-time variance reduction alone lifts the single-sample baseline by +0.069 SSIM and +1.68 dB *without any retraining*, purely by recovering the conditional mean; the dispersion tightens in step (SSIM std 0.106 → 0.084), pointing to fewer catastrophic large-mask failures. Second, the final system (wider backbone, auxiliary reconstruction term, few-step sampling, denser tiles, eight-way flips and Gaussian-weighted blending) reaches 0.862 SSIM, 24.87 dB PSNR and 0.0047 MSE, a further +0.052 SSIM and +3.32 dB. We clear both baselines by a wide margin (+0.107 SSIM, +5.51 dB over the stronger 3D Pix2Pix), although that margin combines the model with the 16-pass inference recipe they do not use. At a matched single-pass budget our single-sample row is already ahead of 3D Pix2Pix on PSNR (+0.51 dB) and MSE while trailing slightly on SSIM (−0.014), so flow matching contributes part of the gain and averaging supplies the rest.

3.3 Training-cohort and held-out performance

The leaderboard covers the validation split only, so Table 2 adds the two splits of the training cohort we can score ourselves under the identical deployed recipe: *training* is a 32-case subset of the volumes the model was optimised on, and *held-out* is the 16-case tail of the same cohort, excluded from every training run

Table 2. Performance of the final system across the three splits under one identical inference recipe (mean ± std). Training and held-out cases are scored locally with our own implementation of the challenge metrics; the validation row is the official leaderboard score.

Split	Cases	SSIM↑	PSNR↑	MSE↓
Training (subset, seen)	32	0.855 ± 0.087	26.54 ± 5.14	0.0041 ± 0.0046
Held-out (unseen)	16	0.885 ± 0.076	27.36 ± 4.93	0.0032 ± 0.0029
Validation (official)	219	0.862 ± 0.090	24.87 ± 4.55	0.0047 ± 0.0035

and used for all internal A/B decisions. The aggregate means are not directly comparable, since the mask-size mix differs (20 of 32 training cases fall in the large-mask bucket $n_h \geq 20$k, against 6 of 16 held-out cases); within each bucket the unseen split scores at or above the seen one (large-mask SSIM 0.814 vs 0.807, small-mask 0.964 vs 0.949), so the network shows no sign of memorising its training volumes.

3.4 Ablations

Table 3 isolates each component along the deployment path. The first four rows use the initial width-32 model with 25-step sampling. Adding overlap-averaged 128^3 tiling to the single-sample baseline gives the largest single jump (+0.059 SSIM / +1.22 dB). Flip augmentation adds a further +0.14 dB by cancelling orientation-dependent bias, and the five-sample ensemble +0.32 dB on top by cancelling the stochastic variance of the ODE; acting on different variance sources, their gains compound. Retraining with the auxiliary SSIM term is worth another +0.017 SSIM. The last two rows move to the deployed recipe. Fine-tuning the sampler for few-step generation, with denser 96^3/stride-32 tiles, eight-way flips and a two-sample ensemble ($2 \times 8 = 16$ passes), is the largest remaining jump (+0.033 SSIM / +2.90 dB); it also cuts the per-patch cost from 25 evaluations to 4, which is what makes the 16-pass budget affordable. Widening the backbone to width-48 and replacing uniform tile averaging with Gaussian blending then adds only +0.002 SSIM / +0.08 dB. These two changes were never run separately, so their individual contributions remain unresolved; the split does show that the final gain is dominated by the sampling recipe rather than by added capacity.

3.5 Qualitative results

Figure 3 shows our fills on four held-out cases from the large-mask regime, where the metric penalty is concentrated. The fills recover a globally consistent tissue layout and continuous brain boundaries, and their intensity and contrast blend seamlessly into the surrounding anatomy at the mask border. Fine texture inside the largest cavities is smoother than the ground truth: the expected signature of the conditional-mean estimator that averaging drives the fill towards (Sec. 4).

Table 3. Ablation along the deployment path on the official BraTS 2026 Local-Inpainting validation set (mean ± std). Each row is a distinct challenge submission. Rows 1–4 add one inference component at a time to the initial width-32 model with 25-step sampling, switching from a single 96^3 patch to overlap-averaged 128^3 tiling at row 2. Row 5 retrains with the auxiliary term and moves from a five- to a three-sample ensemble. Row 6 adds few-step fine-tuning (4 evaluations per patch) with 96^3 stride-32 tiles, a two-sample ensemble, and eight-way flips; row 7 widens the backbone to width-48 and switches uniform tile averaging to Gaussian blending, two changes not run separately.

Configuration	SSIM ↑	PSNR ↑	MSE ↓
Single-sample baseline (96^3)	0.741 ± 0.106	19.87 ± 3.11	0.013 ± 0.008
+ overlap-averaged tiling (128^3)	0.800 ± 0.088	21.09 ± 3.30	0.010 ± 0.007
+ axial flip TTA	0.804 ± 0.087	21.23 ± 3.31	0.010 ± 0.006
+ 5-sample ensemble	0.810 ± 0.084	21.55 ± 3.43	0.009 ± 0.006
+ auxiliary SSIM term (3-sample)	0.827 ± 0.077	21.89 ± 3.45	0.008 ± 0.006
+ few-step sampling & tiles	0.860 ± 0.091	24.79 ± 4.53	0.0048 ± 0.0036
+ wider backbone (**final**)	$\mathbf{0.862 \pm 0.090}$	$\mathbf{24.87 \pm 4.55}$	$\mathbf{0.0047 \pm 0.0035}$

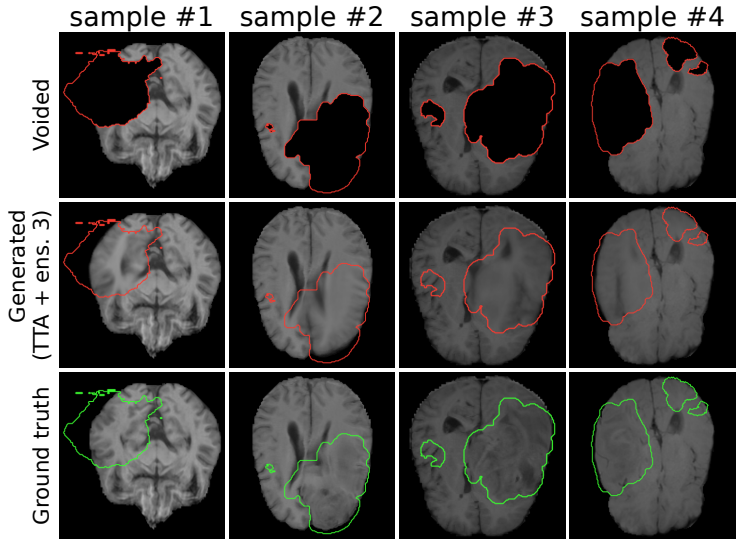

Fig. 3. Held-out reconstructions. Four held-out cases (columns), each an axial slice through the mask centroid, using axial-flip TTA and a three-sample ensemble. *Top:* voided input. *Middle:* our fill. *Bottom:* ground truth. The mask is outlined in red, and in green on the ground-truth row. These are the largest-mask held-out subjects, where reconstruction is hardest.

Figure 4 unrolls the sampling process for one case: the fill begins as noise inside the mask and resolves into coherent anatomy matching the ground-truth structure.

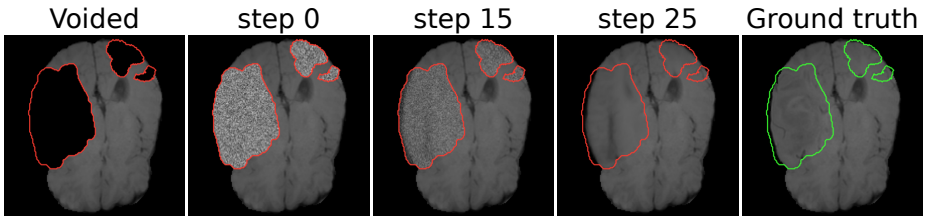

Fig. 4. Sampling trajectory. The composited volume for one held-out case at steps $0, 15, 25$ of the 25-step schedule of the initial model, flanked by the voided input (left) and ground truth (right). The mask is outlined in red, and in green on the ground-truth panel.

4 Discussion

On the cases we inspect qualitatively (Fig. 3), the model produces plausible, spatially coherent completions whose tissue layout and intensity statistics agree with the surrounding anatomy, including well inside large cavities. This is an observation about the outputs shown and the aggregate scores: we ran no targeted analysis (probing the learned representation, or ablating the available boundary context) that would establish *how* the network arrives at those fills, so we make no claim about the prior it has internalised. The ablation splits the total +0.121 SSIM gain into two levers: inference-time sampling alone contributes +0.069 without touching the weights, and the remaining +0.052 comes from joint train- and inference-side changes it attributes mostly to the sampling recipe rather than to added capacity. Those inference gains transfer to any checkpoint, which we see as a strength, but they also mean the headline number reflects the full recipe as much as the model.

Sharpness and distortion also pull against each other: MSE and PSNR provably reward the average of plausible fills rather than a sharp one, and SSIM followed them in every configuration we measured, so the method is tuned to this protocol and a human rater or a texture-sensitive tool might prefer a sharper output. A second limitation is a train/deploy mismatch: training patches are centred on the known healthy region, whereas at test time the model fills the whole lesion mask blindly, which tiled inference only partly mitigates. Variance reduction also costs a 16× longer inference, and our study rests on a single modality, a one-GPU budget, and baseline comparisons we are still extending.

These gaps point to concrete next steps. A wavelet-domain flow would let the model see the full volume at once instead of tiling [7], easing the mismatch above, while a subband-weighted loss, or trained classifier-free guidance [9] on our auxiliary term, would make the sharpness-versus-score trade-off an explicit knob. Longer term, seeding the flow from a learned, condition-dependent start rather than plain noise is the most promising way to close the large-mask gap.

Acknowledgement. Data used in this publication were obtained as part of the Challenge project through Synapse ID (syn74274097). All reported numbers were submitted to the BraTS 2026 Local Inpainting leaderboard by Synapse team **startorch_academy**. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00256629) grant funded by the Korea government (MSIT), and by the ITRC (Information Technology Research Center) support program (IITP-2026-RS-2024-00437718) supervised by IITP.

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

References

1. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: International Conference on Learning Representations (ICLR) (2023)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv preprint arXiv:2107.02314 (2021). https://doi.org/10.48550/arXiv.2107.02314
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.KLXWJJ1Q
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. The Cancer Imaging Archive (2017). https://doi.org/10.7937/K9/TCIA.2017.GJQ7R0EF
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Nature Scientific Data **4**, 170117 (2017). https://doi.org/10.1038/sdata.2017.117
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2016)
7. Danese, D., Lombardi, A., Attimonelli, M., Fasano, G., Di Noia, T.: FlowLet: Conditional 3d brain MRI synthesis using wavelet flow matching. Medical Image Analysis **113**, 104161 (2026). https://doi.org/10.1016/j.media.2026.104161
8. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
9. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: NeurIPS Workshop on Deep Generative Models and Downstream Applications (2021)
10. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2017). https://doi.org/10.1109/CVPR.2017.632

11. Kingma, D.P., Welling, M.: Auto-encoding variational Bayes. In: International Conference on Learning Representations (ICLR) (2014)
12. Kofler, F., Meissen, F., Steinbauer, F., Graf, R., Oswald, E., de la Rosa, E., Paetzold, J.C., Wiestler, B., et al.: The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting. arXiv preprint arXiv:2305.08992 (2023). https://doi.org/10.48550/arXiv.2305.08992
13. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (ICLR) (2023)
14. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (ICLR) (2023)
15. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: RePaint: Inpainting using denoising diffusion probabilistic models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
16. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
17. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
18. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention (MICCAI) (2015)
19. Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH (2022)
20. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021)
21. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861