Now You Have My Healthy Attention: A U-DiT for Brain-MRI Inpainting

Danilo Danese, Angela Lombardi, and Tommaso Di Noia

Politecnico di Bari, Bari, Italy
{danilo.danese, angela.lombardi, tommaso.dinoia}@poliba.it

Abstract. The ASNR-MICCAI BraTS Local Synthesis (Inpainting) task asks for the anatomically plausible completion of healthy brain tissue within a masked region of a T1-weighted MRI, providing a tumor-free anatomical reference for downstream analysis. As the task is scored by distortion metrics (SSIM, PSNR, MSE), we build a deterministic regression model and focus on giving it inductive biases tailored to inpainting. Our network follows the U-DiT principle of performing self-attention on a downsampled token grid: a volumetric encoder-decoder imports long-range context through a downsampled global self-attention block with three-dimensional rotary position embeddings, while convolutions and skip connections preserve high-frequency detail. Two ideas drive our results. First, we constrain the attention so that occluded ("void") tokens attend only to known-healthy tokens of the same volume, with a learned bias toward each query's contralateral homologue, forcing the completion to be inferred from observed anatomy rather than from other unknown regions. Second, we add a contralateral-symmetry input that supplies the mirrored healthy hemisphere as a patient-specific prior; since the brain is approximately bilaterally symmetric and lesions are typically unilateral, this prior improves the distortion metrics at matched structural similarity. On the official BraTS-2026 validation leaderboard our submission reaches a mean healthy-region SSIM of 0.864, PSNR of 24.7 dB and MSE of 4.6×10^{-3} over 219 cases. We further analyse the residual smoothness inherent to distortion-optimal regression and discuss its implications for anatomical realism.

Keywords: Inpainting · BraTS 2026 · MRI · Transformers

1 Introduction

Quantitative analysis of brain tumours from magnetic resonance imaging (MRI) supports diagnosis, surgical planning and treatment monitoring. A recurring obstacle is that a patient-specific healthy reference essentially never exists: the first scan is acquired only after symptom onset, so the anatomical baseline is already compromised. This absence injects a "pathology bias" into registration, atlas construction and segmentation pipelines. Synthesising the missing healthy anatomy in the lesioned region reconstructs a plausible healthy anatomical reference and enables downstream analyses that would otherwise be confounded by

the lesion. The ASNR-MICCAI BraTS *Local Synthesis of Healthy Brain Tissue via Inpainting* task formalises this problem [2]: given a T1-weighted volume with a masked region, a model must inpaint plausible healthy tissue, evaluated over the healthy portion of the mask by the Structural Similarity Index (SSIM) [23], peak signal-to-noise ratio (PSNR) and mean-squared error (MSE).

Prior BraTS inpainting methods fall into two main families. Regression models trained with pixel-wise and SSIM losses [25–27] predict a single completion and achieve strong performance on distortion-based metrics. A recurring ingredient is aggressive random-mask augmentation, which exposes the network to diverse mask shapes and locations during training [26]. In contrast, generative approaches, particularly diffusion models [9–11], sample from the conditional distribution and generate more realistic textures, but typically sacrifice distortion metrics in favour of perceptual quality. We therefore focus on the regression setting and investigate which inpainting-specific architectural priors, rather than generic image-to-image backbones, most effectively improve reconstruction quality. Recent diffusion transformers replace convolutional feature extraction with self-attention over patch tokens [17] and have shown promising results for 3D brain MRI synthesis, including wavelet-domain flow-matching frameworks [8]. Among them, U-DiT improves computational efficiency by applying attention on a downsampled token grid while preserving most of its representational power [21]. Finally, we leverage the approximate bilateral symmetry of the brain, a well-established prior for lesion analysis [15], both by providing the contralateral mirrored hemisphere as an additional input channel and by applying mirror test-time augmentation during inference [22].

Because these metrics are optimised by the conditional mean of plausible completions, we adopt a deterministic regression model and focus on the inductive biases most relevant to image inpainting. We build on the U-DiT idea of performing self-attention on a downsampled token grid [21], which brings global context to a volumetric network at a fraction of the cost of full-resolution attention, and ask what additional structure the inpainting task itself demands. We identify two key inductive biases for this task: reconstruction should rely only on observed healthy tissue, and the contralateral hemisphere provides a strong patient-specific anatomical prior.

Our contributions are:

- A **volumetric network based on the U-DiT architecture**: a downsampled global self-attention block with three-dimensional rotary position embeddings [20, 21] captures long-range contextual dependencies, while convolutions and skip connections carry the local anatomical detail.
- A **non-local "healthy-only" attention with contralateral weighting** that removes occluded tokens from the set of attendable keys, so a query in the void attends exclusively to known-healthy tokens of the same volume, among which a learned, zero-initialised bias peaked at the query's mirror position decides which to read first. This makes the reconstruction draw on observed anatomy rather than on other, unknown regions.

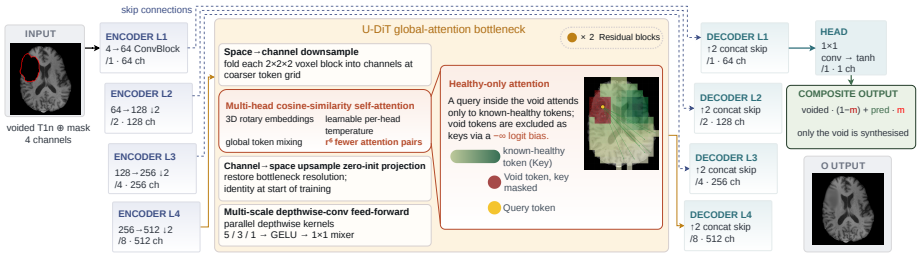

Fig. 1: The U-DiT-style volumetric network. A convolutional encoder–decoder with skip connections carries local detail; the bottleneck holds a single downsampled global self-attention block (space-to-channel $\downarrow r$, cosine-similarity attention with 3D RoPE, channel-to-space $\uparrow r$).

– A **contralateral-symmetry input** that feeds the network the left-right mirror of the voided volume about the estimated mid-sagittal plane, together with a validity channel that discounts mirror voxels that fall outside the brain or inside the void. The contralateral healthy hemisphere provides a strong patient-specific anatomical prior for reconstructing the missing tissue.

Code is available at https://github.com/sisinflab/WaveUDiT.

2 Methods

2.1 Dataset and pre-processing

We use the BraTS-Local-Inpainting data [2], derived from the BraTS-GLI T1-weighted collection [3–6,16] (1,251 training cases). Each case provides the ground-truth T1n, a voided image with a masked region removed and the binary mask; scoring is restricted to the healthy sub-region of the mask. We normalise intensities by the per-volume maximum to $[0, 1]$, rescale to $[-1, 1]$, and crop a $144\times208\times208$ window around the void for training. At inference the known tissue outside the mask is copied back verbatim, so only the void is synthesised.

2.2 Network architecture

The network is a 3D convolutional encoder-decoder (base width 64, four resolution levels, GroupNorm/SiLU) that injects global context through a single U-DiT block [21] at the bottleneck (Fig. 1). The U-DiT principle is to run self-attention not at full resolution but on a downsampled token grid, which yields long-range token mixing at a small fraction of the cost of full-resolution attention while the surrounding convolutions and skip connections preserve high-frequency detail. Concretely, we apply a lossless space-to-channel reshape, the 3D analogue of pixel-unshuffle, that stacks each $r\times r\times r$ block of neighbouring voxels (r=2)

into the channel dimension: a C-channel grid of size $D\times H\times W$ becomes an r^3C-channel grid of size $\frac{D}{r}\times\frac{H}{r}\times\frac{W}{r}$ with no loss of information. The token grid, therefore, holds r^3=8× fewer positions, and since self-attention compares every pair of tokens its cost drops by r^6=64×. A 1×1 convolution then maps the widened channels to the token dimension, multi-head cosine-similarity attention with a learnable per-head temperature and 3D rotary position embeddings [20] mixes the tokens globally, and a channel-to-space reshape (the inverse pixel-shuffle) restores the resolution. The attention output projection is zero-initialised, so the block is an identity at initialisation and the network begins as its convolutional counterpart. Every attention operation in the model is this single coarse-scale block, which supplies exactly the non-local context a convolutional receptive field lacks. The network predicts the full volume with a tanh head, composited as

$$\hat{x} \;=\; x_v \odot (1-m) \;+\; r_\theta(x_v, m) \odot m, \tag{1}$$

with x_v the voided input, m the binary mask (one inside the void, zero on known tissue), and r_θ the network, so known tissue is preserved exactly.

Design rationale: why not a full transformer. Our architecture builds on a wavelet-domain diffusion transformer for 3D brain MRI synthesis [8], based on the hierarchical hourglass transformer [7]. Rather than adopting the full transformer backbone, we retain only its downsampled-attention primitive and integrate it into a convolutional hierarchy, which we found better suited to brain MRI inpainting. In preliminary experiments, a full-transformer backbone plateaued around 0.58 SSIM on our internal split, well below the convolutional model, at substantially higher compute (Section 3). This design is consistent with the nature of the task. The non-local information required for inpainting, such as global anatomy and contralateral context, is predominantly low-frequency and can therefore be captured by a single global-attention block at the bottleneck. In contrast, the convolutional hierarchy preserves fine anatomical detail and provides a stronger inductive bias than a full transformer when training on the available cases with a distortion-based objective.

Tokenizer. The linear bottleneck patchify is augmented with a residual overlapping-convolution branch: a band-grouped convolution for the token merge and a depthwise convolution before a pixel-shuffle [19] for the split, giving each token cross-boundary spatial context. The branch output projection is zero-initialised, so it is inactive at initialisation and preserves the identity start.

2.3 Non-local healthy-only attention with contralateral weighting

Left unconstrained, the bottleneck attention lets a query inside the void pool information from other void tokens, content that is itself unknown and being predicted, which turns reconstruction into a fixed-point over missing values. We remove this coupling, and at the same time tell each query which of the remaining

tokens to prefer, through a single additive term on the attention logits. For a query token i and key token j,

$$\ell_{ij} = s\,\frac{q_i^\top k_j}{\|q_i\|\,\|k_j\|} + b_j + \lambda_h\,\bar{v}_j\,\exp\!\Big(-\tfrac{\|p_j-\mu(p_i)\|^2}{2\sigma^2}\Big), \tag{2}$$

with a learnable per-head temperature s, normalised into attention weights $\alpha_{ij} = \mathrm{softmax}_j(\ell_{ij})$. The bias b_j is read off the mask m average-pooled onto the coarse token grid, each token aggregating the r^3 voxels of its own block: writing $\bar{m}_j$ for the resulting per-token void fraction, $b_j = -\infty$ where $\bar{m}_j > 0.5$ and $b_j = 0$ elsewhere, so every predominantly-void key receives zero weight and each query, void or observed, attends only to known-healthy tokens of the same volume, reaching non-locally across both hemispheres (Fig. 2). A guard disables the bias for the degenerate all-void sample to avoid an undefined softmax. This part is parameter-free and, as Section 3 shows, the most effective single change we make.

The last term then decides which admissible key to read first, since the healthy tissue that best predicts a void voxel is usually its own contralateral homologue [15]: p_i is the grid position of token i, $\mu(p_i)$ its reflection about the estimated mid-sagittal plane, and the per-head strength λ_h and width σ are learned. Three properties make it safe to add to an already-trained network. The plane is the same per-volume estimate used for the contralateral input channel rather than the centre of the token grid, so the preference follows the anatomy and not the field of view. The gate $\bar{v}_j$ is the pooled mirror-validity map, which suppresses the term exactly where the mirror falls outside the brain or inside the void, as happens for midline-crossing or bilateral lesions.

2.4 Contralateral-symmetry input

We condition the network on the brain's bilateral symmetry [15] by supplying two additional input channels alongside the voided volume x_v and mask m. We first estimate the mid-sagittal plane per sample as the intensity-weighted centroid along the left–right axis h,

$$c = \frac{\sum_h h\,w_h}{\sum_h w_h}, \qquad w = (x_v + 1)/2, \tag{3}$$

where the weight w rescales the normalised intensities to $[0, 1]$ so background voxels contribute nothing. Reflecting the left-right coordinate about c gives the mirror index $h' = 2c - h$ (rounded to the nearest voxel), where the axis runs over $h \in \{0, \ldots, H-1\}$. The mirrored volume copies the voided intensity from that index, $\tilde{x}_h = x_v[\,h'\,]$, and a per-voxel validity map verifies whether the mirror source is usable:

$$\bar{v}_h = \begin{cases} 1 & \text{if } 0 \le h' < H \text{ and } m[\,h'\,] = 0, \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

that is, $\bar{v}_h = 1$ only where the mirror index stays inside the volume and its source is observed tissue rather than void ($m = 1$ marks the masked, to-be-synthesised

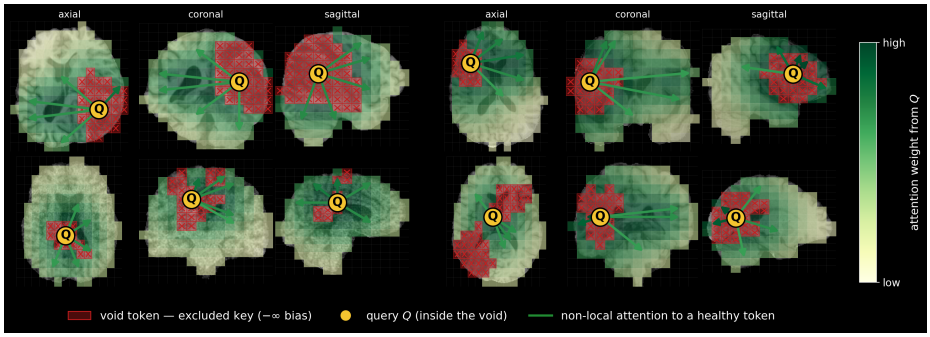

Fig. 2: Healthy-only attention with contralateral weighting, shown for four validation cases (rows), each in three anatomical views (columns). Void tokens (red ×) receive a $-\infty$ key bias and are therefore excluded as keys, so a query Q inside the void (gold) reads only known-healthy tokens (green, shade = attention weight) and can reach them non-locally across the whole volume. In the axial and coronal views, where the left-right axis lies in the plane, the learned bias visibly concentrates the weight on the homologous region of the opposite hemisphere; the sagittal view shows the healthy-only weighting alone, since there the homologue falls outside the displayed slice.

voxels). Conditioning is obtained by concatenation, so the network receives the four-channel input $[\,x_v, m, \tilde{x}, \bar{v}\,]$ and the composited completion becomes:

$$\hat{x} = x_v \odot (1 - m) + r_\theta(x_v, m, \tilde{x}, \bar{v}) \odot m, \tag{5}$$

here $\tilde{x}$ supplies a candidate value for each masked voxel and $\bar{v}$ marks where that value is defined. Because homologous structures across the mid-line are nearly identical in healthy anatomy and lesions are usually unilateral [15], $\tilde{x}$ provides a template for tissue missing on one side (Fig. 3), while $\bar{v}$ lets the network discount invalid (out-of-brain or itself-voided) mirror content, for example in bilateral or mid-line lesions where the mirror source is unavailable. The stem weights on the two extra channels are zero-initialised, so the model reproduces its no-contralateral counterpart and can be warm-started from it.

2.5 Training objective and schedule

We minimise a weighted sum of masked error terms over the void, plus a high-pass term that discourages over-smoothing:

$$\mathcal{L} = \lambda_1 \frac{\sum |\hat{x}-x| \odot w}{\sum w} + \lambda_2 \frac{\sum (\hat{x}-x)^2 \odot w}{\sum w} + \lambda_s \big(1 - \mathrm{SSIM}_m(\hat{x}, x)\big) + \lambda_{\mathrm{hf}}\, \mathcal{L}_{\mathrm{hf}}, \tag{6}$$

with λ_1=1, λ_2=10, λ_s=1 and λ_{hf}=0.5. The squared-error term dominates because the official per-case PSNR rank is exactly the inverse MSE rank, so the

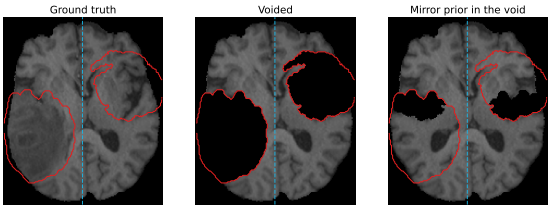

Fig. 3: Contralateral-symmetry input on a double-lesion case with two voids in opposite hemispheres. **Left:** the ground-truth volume, with the two void regions outlined in red and the estimated mid-sagittal plane dashed. **Middle:** the voided input the network actually receives. **Right:** filling each void with its contralateral mirror. Because the two lesions are nearly contralateral, part of one void mirrors onto the other (or onto out-of-brain background) and is left unfilled; the validity channel (not shown) flags exactly these voxels so the network learns to ignore them, while the remaining void is given a plausible anatomical template.

score is two parts MSE to one part SSIM. The two error terms use the weight map:

$$w = m \odot \big(1 - (1-\gamma)\,\mathbf{1}[\text{tumour}]\big), \tag{7}$$

with γ=0.25, which down-weights the void voxels that lie on tumour tissue and are therefore never scored, whereas the SSIM term keeps the full void m because it is a windowed statistic that needs the surrounding context. The high-pass term is the masked L_1 distance between the high-frequency residuals of prediction and target,

$$\mathcal{L}_{\text{hf}} = \frac{\sum |\text{HF}(\hat{x}) - \text{HF}(x)| \odot m}{\sum m}, \tag{8}$$

with

$$\text{HF}(x) = x - G_{\sigma_{\text{hf}}} * x \tag{9}$$

and $G_{\sigma_{\text{hf}}}$ a Gaussian of σ_{hf}=1 voxel; unlike the other terms it is not minimised by blurring, since smoothing the prediction drives $\text{HF}(\hat{x})$ towards zero and widens the gap. Augmentation follows masked image modelling [12]: instead of always re-using the provided void, at each step we draw a real mask from a bank of inpainting masks (eight per case per epoch) and crop a random window around it, so the network sees many void shapes, sizes and positions rather than memorising a single. The bank contains whole real BraTS inpainting masks, including multi-component and peri-lesional regions; masks are not split into connected components and are randomly placed within the brain. We optimise with AdamW [14] at batch size 2 with zero weight decay. We keep an EMA of the weights [18] (EMA decay 0.995) for evaluation and for model selection on the mean healthy-region SSIM over 60 held-out training cases. The training ran for up to 400 epochs with 200 steps warmup and cosine schedule peaking at 10^{-4} and annealed to 10^{-7} [13].

Table 1: Top: ablation on our validation split. Middle: final architectures scored with the official metric on the native volume, mean (std) over the 60 held-out cases. Bottom: official validation leaderboard over 219 cases, mean (std).

Configuration	SSIM↑	PSNR↑	MSE↓
Full transformer [8]	0.580	–	–
U-DiT backbone	0.848	21.5	0.0053
+ healthy-only attention	0.853	21.9	0.0049
+ contralateral input	0.856	22.1	0.0047
+ TTA & annealing (no contralateral attention)	0.865	22.5	0.0044
U-DiT (ours), official metric	**0.880** (0.094)	**23.83** (2.58)	**0.0043** (0.0034)
Wavelet U-DiT	0.870	23.1	0.0048
Wavelet HDiT [8]	0.781	17.9	0.0142
Leaderboard, 219 cases	**0.864** (0.085)	**24.71** (4.03)	**0.0046** (0.0032)

2.6 Inference

At test time we apply mirror test-time augmentation [22], averaging the prediction with its sagittal flip. Inference is a single deterministic forward pass per view (two with the flip). The composited prediction is de-normalised and written back into the original volume geometry.

3 Experiments and Results

3.1 Evaluation metrics and protocol

The task is scored on the healthy sub-region of the mask by three distortion metrics, SSIM, PSNR and MSE, and the headline result is the score returned by the official BraTS-2026 evaluation server. For the ablation we additionally report relative comparisons on our own held-out validation split of the training data, using EMA weights and the official healthy-region convention.

3.2 Ablation study

Table 1 isolates the effect of each design choice on the internal split, added cumulatively. Replacing the full-transformer (hierarchical hourglass with neighbourhood attention) backbone by our convolutional U-DiT hierarchy is decisive: the full transformer plateaus around 0.58 SSIM, whereas the U-DiT backbone alone reaches 0.848. Restricting the bottleneck attention to known-healthy tokens then improves all three metrics and is our single most effective change; adding the contralateral-symmetry input yields a further, smaller gain concentrated on the distortion metrics (PSNR, MSE) at essentially unchanged SSIM, consistent with its role as a low-frequency anatomical template; and mirror test-time augmentation together with the annealed learning-rate schedule contributes

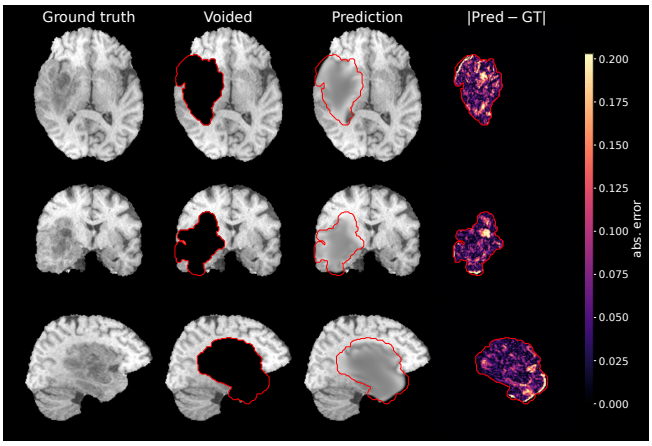

Fig. 4: Qualitative completion for a validation case (our submitted mirror-TTA model). Columns, left to right: ground truth, voided input with the void boundary outlined in red, our completion and the absolute error $|\hat{x} - x|$. Rows: axial, coronal and sagittal views through the void centroid.

a final increment. We also evaluated two wavelet-domain variants [8]: moving our backbone onto a 3D wavelet decomposition processes eight times fewer spatial positions, and replacing its bottleneck with an hourglass transformer reduces the cost further. Both are cheaper to run, but neither matches the full-resolution U-DiT on the official metric, even though the wavelet variant recovers slightly more anatomical structure inside the void (Fig. 5), so we retain the U-DiT for the submission. We also ablated three additions that did not improve the official distortion score and were therefore excluded from the submitted model: averaging our output with a separate CNN inpainter, matching the hole mean to the surrounding tissue shell, and greedy weight-space souping [24] of our (strongly correlated) checkpoints. The submitted model is thus a single network with mirror test-time augmentation and contralateral attention.

3.3 Validation leaderboard

We uploaded our predictions to the official BraTS-2026 Task-4 (Inpainting) validation queue on the challenge platform [1]. Table 1 reports the scores it returns over the 219-case validation set. The official scores differ from our internal split in both directions, with lower SSIM and higher PSNR, which is expected given the different case distributions. Figure 4 shows a representative completion: known tissue is preserved exactly, so the error is supported only inside the void.

4 Discussion

Our results are driven by matching the inductive bias of the network to the structure of the inpainting task. *Where* global context is injected matters more

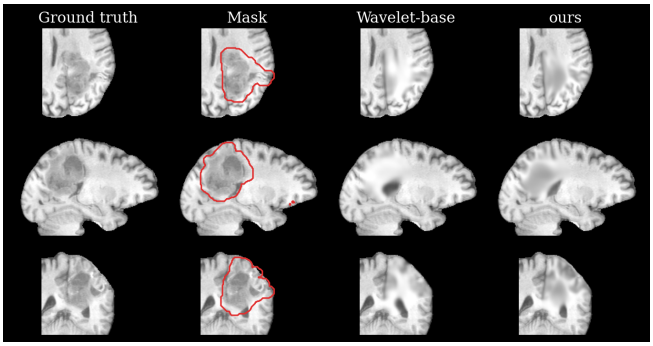

Fig. 5: Perception-distortion trade-off on one case, three views. Columns, left to right: ground truth, the void mask (red), our wavelet variant and our metric-optimal completion (U-DiT). The wavelet variant recovers slightly more anatomical structure inside the void yet scores lower on all three distortion metrics, so the texture a reader judges more realistic is not what the score rewards.

than simply adding it: an unconstrained attention bottleneck moves the score little, whereas forcing occluded tokens to read only known-healthy context makes reconstruction an inference from observed anatomy rather than a fixed-point over unknown regions. This change is parameter-free and applies to any inpainting network with a coarse attention stage, which makes it a broadly reusable design. The contralateral-symmetry input is complementary: it hands the network an explicit, patient-specific template for the missing side and improves the distortion metrics at matched structural similarity. Its benefit persists alongside mirror test-time augmentation even though both exploit the brain's left-right symmetry, indicating that a symmetry prior is more effective when supplied to the model as conditioning than when applied only as a post-hoc average.

Realism versus the distortion metric. Like other distortion-optimised regression models, our completions remain smooth. Since SSIM, PSNR, and MSE are optimised by the conditional mean of plausible completions, predictions tend to suppress uncertain high-frequency details while preserving low-frequency anatomical structure. Consequently, the synthesised tissue contains only a fraction of the high-frequency texture energy of real tissue: measured over the scored region with the same high-pass operator the training term uses, the completions retain 0.65 ± 0.11 of its high-frequency RMS. This is the expected behaviour on a single-reference distortion score rather than a failure of the model, but it does mean the metric and anatomical realism can diverge (Fig. 5). Closing this gap while preserving the distortion score, for instance with a residual generative refinement stage that adds high-frequency detail without disturbing the low-frequency content the metric rewards, is the most promising direction for future work, and motivates complementing SSIM with texture-aware measures.

5 Conclusion

We presented a U-DiT-style volumetric network for healthy brain-tissue inpainting, in which a single downsampled global self-attention block with 3D rotary position embeddings supplies long-range context while convolutions and skip connections preserve high-frequency detail. Two task-specific ideas drive its performance: restricting the attention so that occluded tokens read only known-healthy context, and a contralateral-symmetry input that gives the network the patient's own healthy hemisphere as a template for the missing tissue. On the official BraTS-2026 validation leaderboard the submission reaches a mean healthy-region SSIM of 0.864, PSNR of 24.71 dB and MSE of 4.57×10^{-3} over 219 cases. Our analysis of the residual smoothness of distortion-optimal regression points to residual generative refinement as the route to greater anatomical realism without sacrificing the distortion metrics.

Acknowledgements. We thank the BraTS challenge organisers for the data and evaluation platform.
Synapse team: **Morpheus**.

References

1. et al., A.K.: Federated benchmarking of medical artificial intelligence with medperf. Nat. Mac. Intell. **5**(7), 799–810 (2023)
2. et al., F.K.: The brain tumor segmentation (brats) challenge: Local synthesis of healthy brain tissue via inpainting (2024), https://arxiv.org/abs/2305.08992
3. et al., U.B.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021), https://arxiv.org/abs/2107.02314
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: 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
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: 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
6. 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. Scientific Data **4**(1), 170117 (Sep 2017). https://doi.org/10.1038/sdata.2017.117, https://doi.org/10.1038/sdata.2017.117
7. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In: ICML. Proceedings of Machine Learning Research, vol. 235, pp. 9550–9575. PMLR / OpenReview.net (2024)
8. Danese, D., Lombardi, A., Fasano, G., Attimonelli, M., Noia, T.D.: Wavedit: Distribution-aware wavelet flow matching for efficient 3d brain MRI synthesis. CoRR **abs/2606.08670** (2026)

9. Durrer, A., Bieder, F., Friedrich, P., Menze, B.H., Cattin, P.C., Kofler, F.: fastwdm3d: Fast and accurate 3d healthy tissue inpainting. In: DGM4MICCAI@MICCAI. Lecture Notes in Computer Science, vol. 16128, pp. 171–181. Springer (2025)
10. Durrer, A., Wolleb, J., Bieder, F., Friedrich, P., Melie-García, L., Ocampo-Pineda, M., Bercea, C.I., Hamamci, I.E., Wiestler, B., Piraud, M., Yaldizli, Ö., Granziera, C., Menze, B.H., Cattin, P.C., Kofler, F.: Denoising diffusion models for 3d healthy brain tissue inpainting. In: DGM4MICCAI@MICCAI. Lecture Notes in Computer Science, vol. 15224, pp. 87–97. Springer (2024)
11. Ferreira, A., Luijten, G., Puladi, B., Kleesiek, J., Alves, V., Egger, J.: Brain tumour removing and missing modality generation using 3d WDM. CoRR **abs/2411.04630** (2024)
12. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: CVPR. pp. 15979–15988. IEEE (2022)
13. Loshchilov, I., Hutter, F.: SGDR: stochastic gradient descent with warm restarts. In: ICLR (Poster). OpenReview.net (2017)
14. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (Poster). OpenReview.net (2019)
15. Ma, Y., Wang, D., Liu, P., Masters, L., Barnett, M., Cai, T.W., Wang, C.: Symmetry awareness encoded deep learning framework for brain imaging analysis. In: MICCAI (12). Lecture Notes in Computer Science, vol. 15012, pp. 742–752. Springer (2024)
16. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahaniy, K., Kirby, J., Burren, Y., Porz, N., Slotboomy, J., Wiest, R., Lancziy, L., Gerstnery, E., Webery, M.A., Arbel, T., Avants, B., Ayache, N., Buendia, P., Collins, L., Cordier, N., Van Leemput, K.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Transactions on Medical Imaging **99** (12 2014). https://doi.org/10.1109/TMI.2014.2377694
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV. pp. 4172–4182. IEEE (2023)
18. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization **30**(4), 838–855 (1992). https://doi.org/10.1137/0330046, https://doi.org/10.1137/0330046
19. Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: CVPR. pp. 1874–1883. IEEE Computer Society (2016)
20. Su, J., Ahmed, M.H.M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
21. Tian, Y., Tu, Z., Chen, H., Hu, J., Xu, C., Wang, Y.: U-dits: Downsample tokens in u-shaped diffusion transformers. In: NeurIPS (2024)
22. Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T.: Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing **338**, 34–45 (2019)
23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
24. Wortsman, M., Ilharco, G., Gadre, S.Y., Roelofs, R., Lopes, R.G., Morcos, A.S., Namkoong, H., Farhadi, A., Carmon, Y., Kornblith, S., Schmidt, L.: Model soups:

averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In: ICML. Proceedings of Machine Learning Research, vol. 162, pp. 23965–23998. PMLR (2022)
25. Zhang, J., Chen, K., Weng, Y.: Synthesis of healthy tissue within tumor area via u-net. In: BraTS/CrossMoDA@MICCAI. Lecture Notes in Computer Science, vol. 14669, pp. 233–240. Springer (2023)
26. Zhang, J., Weng, Y., Chen, K.: Robust 3d brain MRI inpainting with random masking augmentation. In: BraTS-Lighthouse/AIMS-TBI@MICCAI (2). Lecture Notes in Computer Science, vol. 16377, pp. 102–109. Springer (2025), winner, BraTS-Inpainting 2025
27. Zhang, J., Weng, Y., Chen, K.: U-net based healthy 3d brain tissue inpainting. CoRR **abs/2507.18126** (2025), winner, BraTS Local-Synthesis/Inpainting 2024