



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.

WaveDiT: Distribution-Aware Wavelet Flow Matching for Efficient 3D Brain MRI Synthesis

Danilo Danese¹, Angela Lombardi¹, Giuseppe Fasano¹, Matteo Attimonelli^{1,2},
and Tommaso Di Noia¹

¹ Politecnico di Bari, Italy {name.surname}@poliba.it

² Sapienza University of Rome, Italy

Abstract. Large and demographically balanced datasets are essential for reliable neuroimaging biomarkers. Full-resolution 3D brain MRI synthesis can support data augmentation in this setting, but existing approaches either incur prohibitive computational cost at volumetric scale or rely on lossy latent compression that may compromise anatomical detail. As a result, practical 3D generative augmentation often requires specialized compute infrastructure. We propose WaveDiT, a conditional flow matching framework operating in the coefficient space of a 3D Haar Discrete Wavelet Transform. The model combines factorized spatio-depth attention with band-wise heteroscedastic uncertainty modeling derived from higher-order wavelet statistics. Predicted log-variance is integrated directly into both the flow objective and conditioning pathway, enabling adaptive precision consistent with the heavy-tailed and input-dependent variance structure of anatomical detail. This formulation supports full-resolution 3D synthesis under practical memory and time constraints on a single modern GPU. Evaluation on a multi-site cohort demonstrates improved alignment between generated and real MRI distributions, together with enhanced downstream brain age prediction and region-level anatomical agreement relative to diffusion, latent, and wavelet-based baselines. Code is available at <https://github.com/sisinflab/WaveDiT>.

Keywords: Flow Matching · Diffusion Transformers · 3D MRI imaging

1 Introduction

High-resolution brain MRI is fundamental to neuroimaging research, supporting clinically relevant tasks such as brain age prediction (BAP), disease-risk stratification, and longitudinal monitoring of neurodegeneration [5, 25]. Developing robust biomarkers for these tasks typically requires large, demographically balanced cohorts [1]. However, acquiring such datasets is challenging due to high acquisition costs, privacy constraints, and site-specific heterogeneity. This scarcity is particularly pronounced in specific age ranges, where sparsity can induce biased regression and inflated uncertainty in normative aging trajectories [9]. Generative modeling offers a promising path forward, enabling the synthesis of balanced cohorts to augment scarce clinical data [4]. Despite recent progress, scaling

generative models to full-resolution 3D MRI remains computationally demanding. Pixel-space diffusion models achieve high fidelity but require hundreds to thousands of iterative denoising steps over millions of voxels [15], resulting in substantial training and inference costs. Training such models at full volumetric resolution often necessitates high-memory GPUs and prolonged compute schedules, limiting accessibility outside specialized infrastructure. Latent diffusion reduces computational demand through learned compression [26, 24]; however, this compression is inherently lossy and may discard fine-grained anatomical detail or introduce reconstruction artifacts, particularly in regions with subtle cortical structure [22]. The Discrete Wavelet Transform (DWT) provides an alternative representation that preserves invertibility while reducing spatial dimensionality. By decomposing volumes into low-frequency approximation and high-frequency subbands, wavelets maintain anatomical structure without learned compression artifacts. Recent wavelet-based generative approaches demonstrate that operating in the wavelet domain is effective for 3D MRI synthesis [12, 7]. However, existing methods typically treat all subbands uniformly, applying band-agnostic objectives and conditioning strategies. In practice, wavelet subbands exhibit markedly different statistical properties: approximation coefficients remain near-Gaussian, whereas high-frequency bands are sparse, heavy-tailed, and strongly heteroscedastic, with distributions that evolve along the generative trajectory. This heteroscedastic structure implies input-dependent predictive uncertainty, making uniform loss weighting suboptimal. Heteroscedastic uncertainty modeling has improved robustness in regression [17, 27] and medical image registration [34], and its integration within flow-based generative modeling remains limited. To address these challenges, we propose (i) **WaveDiT**, a conditional flow matching framework that operates directly in the wavelet domain. WaveDiT extends the Hourglass Diffusion Transformer (HDiT) [6] to volumes with factorized intra-slice and inter-slice attention, avoiding computational complexity over full 3D self-attention in the wavelet domain. Furthermore, we introduce (ii) **Morpheus**, a state-aware auxiliary network that predicts per-band precision from higher-order wavelet statistics. Morpheus enables a Bayesian heteroscedastic loss and frequency-aware conditioning that adapts to the signal’s complexity along the flow trajectory. We evaluate the proposed framework using a multi-level protocol that combines global distributional metrics with downstream brain age prediction and region-level anatomical analysis.

2 Methods

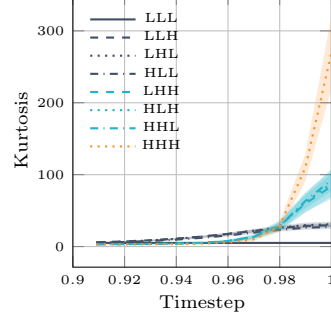
2.1 Wavelet modeling

WaveDiT operates in the coefficient space of a single-level 3D Haar DWT, which decomposes each volume as $\mathcal{W} : \mathbb{R}^{1 \times D \times H \times W} \rightarrow \mathbb{R}^{8 \times D' \times H' \times W'}$, with $D' = D/2$, $H' = H/2$, $W' = W/2$, producing one low-frequency approximation subband (LLL) and seven directional HF detail subbands ($LLH, LHL, \dots, HHH$).

Analysis of the training volumes reveals two key statistical properties. First, the signal energy concentrates heavily in the approximation band: LLL con-

Table 1: Wavelet band statistics and kurtosis evolution along the flow trajectory. **Left:** per-band kurtosis at $t = 1$ and Pearson correlation between local HF variance and LLL intensity. **Right:** subband kurtosis as a function of timesteps.

Band	κ at $t = 1$	Ratio	Pearson r
LLL	5.06 ± 0.87	$1.7\times$	—
LLH	29.77 ± 2.71	$9.9\times$	0.559 ± 0.037
LHL	30.99 ± 3.72	$10.3\times$	0.592 ± 0.040
HLL	27.32 ± 2.88	$9.1\times$	0.466 ± 0.092
LHH	83.80 ± 17.68	$27.9\times$	0.431 ± 0.044
HLH	87.95 ± 18.17	$29.3\times$	0.354 ± 0.068
HHL	92.53 ± 15.36	$30.8\times$	0.399 ± 0.078
HHH	269.44 ± 50.66	$89.8\times$	0.200 ± 0.066



tains 98.11% of total energy. Second, wavelet coefficient distributions change drastically along the flow trajectory (Table 1). At $t = 0$ (pure noise), all bands exhibit Gaussian statistics with kurtosis $\kappa \approx 3$. As the flow progresses toward $t = 1$ (data), the *LLL* band remains near-Gaussian ($\kappa \approx 5$), but HF bands diverge sharply: single-axis subbands develop $\kappa \in [27, 31]$, two-axis subbands reach $\kappa \in [84, 93]$, and the isotropic *HHH* band peaks at $\kappa \approx 270$ a ratio of $89.8\times$ relative to $t = 0$. This evolving distributional contrast, confirmed by Jarque-Bera tests that reject normality at $t = 1$ with $p < 10^{-10}$, makes kurtosis a natural signal-noise discriminator. Beyond distributional shape, HF coefficients are also strongly heteroscedastic. Their local variance varies by roughly eight orders of magnitude across space, increasing near tissue boundaries (steep gradients) and decreasing in homogeneous regions. This variance is input-dependent, with diminishing correlation as frequency increases. This induces a mismatch with homoscedastic flow objectives, which assume a single global precision across bands and spatial locations. This motivates an explicit band-wise uncertainty formulation, where precision is adapted to the evolving statistics of each subband.

2.2 Morpheus: State-Aware Uncertainty Modelling

The heteroscedastic, heavy-tailed statistics have direct implications for optimization. Standard flow matching employs uniform MSE weighting, treating all wavelet bands and spatial locations identically. Such fixed-precision losses over-penalize errors at anatomical boundaries (high variance) and under-penalize homogeneous regions (low variance). WaveDiT addresses this mismatch with Morpheus, a lightweight network that predicts input-dependent band-wise uncertainty and influences both optimization and generation.

Feature Extraction. Unlike conventional schedulers that condition solely on timestep t , Morpheus computes the statistical signature of the current noisy state $\mathbf{x}_t$ of an input $\mathbf{x}$ at time t along the flow trajectory. For each channel c , we extract six statistics:

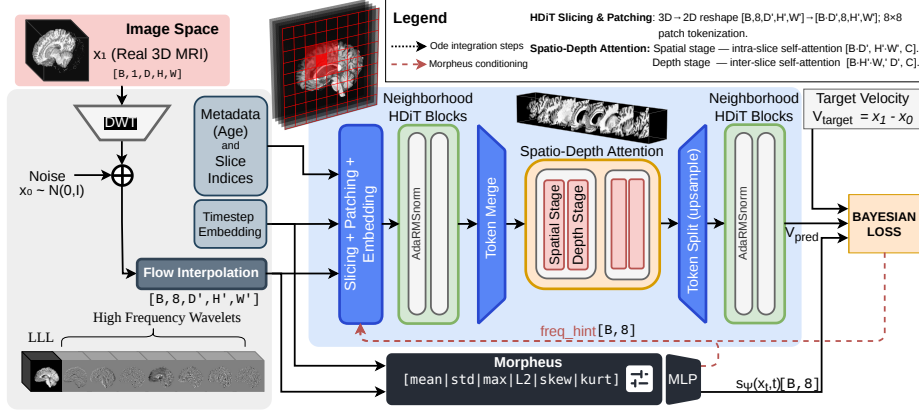


Fig. 1: Training pipeline: wavelet decomposition, HDiT backbone with Morpheus scheduling, and Bayesian heteroscedastic loss.

- **Mean** μ_c and **Standard Deviation** σ_c : First and second moments;
- **Maximum Absolute Value** $\max |x_c|$: Outlier detection;
- **L2 Norm** $\|x_c\|_2 / \sqrt{N}$: Normalized energy per band;
- **Skewness** $\gamma_1 = \mathbb{E}[(x - \mu)^3] / \sigma^3$: Distribution asymmetry;
- **Kurtosis** $\gamma_2 = \mathbb{E}[(x - \mu)^4] / \sigma^4$: Tail heaviness.

These features are concatenated with a sinusoidal time embedding and processed by an MLP (parameters ψ), to produce the band-wise log-variance $\mathbf{s}_\psi(\mathbf{x}_t, t)$.

Bayesian Heteroscedastic Objective. The generative trajectory follows Rectified Conditional Flow Matching [20]. Given $\mathbf{x}_1 \sim p_{\text{data}}$ and $\mathbf{x}_0 \sim \mathcal{N}(0, I)$, the interpolation $\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1$ defines a linear path with target velocity $\mathbf{v}_{\text{target}} = \frac{\mathbf{x}_1 - \mathbf{x}_t}{1 - t + \epsilon}$, with small $\epsilon > 0$ preventing divergence as $t \rightarrow 1$. Following heteroscedastic regression [17, 27], velocity prediction is modeled as $p(\mathbf{v} | \mathbf{v}_\theta, \mathbf{s}) = \mathcal{N}(\mathbf{v}; \mathbf{v}_\theta, e^{\mathbf{s}})$, yielding the objective (backbone θ , Morpheus ψ):

$$\mathcal{L} = \mathbb{E}_{\mathbf{x}_0, \mathbf{x}_1, t} \left[\frac{1}{2} e^{-\mathbf{s}_\psi(\mathbf{x}_t, t)} \|\mathbf{v}_\theta(\mathbf{x}_t, t, \text{cond}) - \mathbf{v}_{\text{target}}\|^2 + \frac{1}{2} \mathbf{s}_\psi(\mathbf{x}_t, t) \right]. \quad (1)$$

This formulation recasts flow matching in wavelet space as a heteroscedastic regression problem aligned with subband statistics. Here, $e^{-\mathbf{s}}$ adaptively reweights the velocity loss during training, down-weighting inherently unpredictable high-frequency content, while $\frac{1}{2}\mathbf{s}$ prevents trivial variance inflation. This results in state-dependent precision, with higher predictive variance during early noisy states and progressively sharper weighting as structured anatomy emerges.

Frequency Conditioning. Morpheus also conditions the backbone. The predicted log-variances $\mathbf{s}$ are linearly projected and concatenated with time, slice, and metadata embeddings to form a *frequency hint*. Active during both training and sampling, this pathway allows the backbone to adapt its predictions to the current reliability of each wavelet band.

2.3 The WaveDiT Architecture

Even in the reduced wavelet domain, global self-attention over the full 3D coefficient tensor remains prohibitive. For a 224^3 input, the DWT yields $D' = H' = W' = 112$, producing $N = D'H'W' \approx 1.4 \times 10^6$ spatial tokens. Full 3D attention scales as $\mathcal{O}(N^2) \approx 2 \times 10^{12}$. WaveDiT avoids this extending the hourglass design [6] to 3D, treating the volume as a batch of 2D slices and restores volumetric coherence through factorized depth attention, reducing complexity to $\mathcal{O}(D'(H'W')^2 + H'W'D'^2) \approx 1.8 \times 10^{10}$, a $\sim 110\times$ reduction.

Slice Processing. The wavelet tensor $\mathbf{w} \in \mathbb{R}^{B \times 8 \times D' \times H' \times W'}$ is reshaped into a batch of $B \cdot D'$ 2D slices and tokenized with 2D patch embeddings. To recover explicit slice position lost in this operation, we encode each slice with fixed random Fourier features, which are combined with sinusoidal time embedding, demographic metadata (age), and the Morpheus frequency hint into a global conditioning vector. All transformer blocks are modulated via AdaRMSNorm [33] as a multiplicative scale, injecting time-, slice-, and frequency-dependent conditioning without concatenating condition vectors to spatial tokens.

Level 1: Neighborhood Attention. At the highest spatial resolution, shallow transformer layers employ 2D sliding-window attention. Each query token attends only to keys within a local $K \times K$ neighborhood, reducing per-layer complexity from $\mathcal{O}((H'W')^2)$ to $\mathcal{O}(H'W'K^2)$. This locality inductive bias aligns naturally with HF coefficients, which encode edges and tissue boundaries with strong local spatial correlation. AxialRoPE [14] is applied to queries and keys, encoding relative 2D positions while preserving translation equivariance.

Level 2: Factorized Spatio-Depth Attention. Deeper layers in the model hierarchy employ our two-stage factorized attention block, which restores volumetric consistency without incurring the cost of full 3D self-attention. In the spatial (intra-slice) stage, we apply global self-attention independently within each 2D slice to capture long-range in-plane dependencies such as bilateral symmetry and ventricle-to-cortex relationships. In the subsequent depth (inter-slice) stage, tokens at each spatial location (h, w) are grouped and attend along the depth axis, enabling cross-slice propagation of anatomical context (e.g., continuity of cortical sulci or the corpus callosum along the Z-axis). This mechanism preserves intra-slice structure and inter-slice consistency at reduced cost.

2.4 Inference

Noise $\mathbf{z}_0 \sim \mathcal{N}(0, I)$ is integrated from $t = 0$ to $t = 1$ via a second-order Heun solver, with Morpheus providing frequency conditioning at each step. The inverse 3D Haar DWT reconstructs the synthesized volume in voxel space.

3 Experiments and Results

Dataset. We follow the multi-site cohort introduced in [7], merging OpenBHB [10], ADNI [23], and OASIS-3 [21] (5,989 cognitively normal subjects, ages 5.9-95.5 yrs)

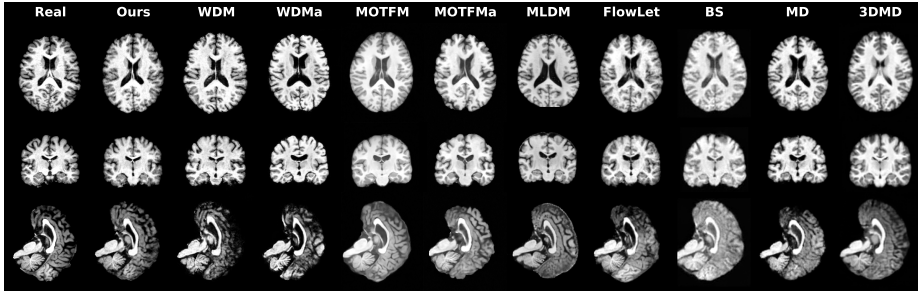


Fig. 2: Visual comparison of models. Axial, coronal, and sagittal views of a real 72 years old subject and age-conditioned generations at the same target age.

to mitigate the elderly under-representation inherent in OpenBHB alone. All volumes undergo standard preprocessing: bias-field correction [30], affine registration to MNI152 [16, 11], skull stripping [28], and $[0, 1]$ normalization, yielding an isotropic $182 \times 218 \times 182$ resolution. A subject-disjoint 20% hold-out is reserved exclusively for BAP evaluation, while all generative models are trained on the last 80% following strict subject-level separation to prevent identity leakage.

Implementation. WaveDiT (patch 8×8 , two HDiT levels: depth 2 each, width 1024, FFN dim 4096) uses level 1 neighborhood attention ($d_{\text{head}} = 64$, window $K = 7$) and level 2 spatio-depth attention ($d_{\text{head}} = 64$), totalising $\sim 142\text{M}$ parameters. The Morpheus MLP has two hidden layers (width 128) with SiLU activations. The model is trained with AdamW (lr= 10^{-4} , weight decay 10^{-2} , bs=4, dropout 0.1) for 200 epochs on a single H100 ($\sim 12\text{GB}$ VRAM at bs=1). Checkpoints are selected by best validation loss; Morpheus is jointly optimized. Training completes in 26 hours on a single H100, compared to ~ 6 days for FlowLet and ~ 7 days for WDM under the same hardware setting. At inference, WaveDiT generates a full 3D volume in ~ 1 second using 10 steps, compared to ~ 6 seconds for 10 steps in FlowLet and ~ 150 seconds for 1000 steps in WDM.

Baselines. We compare nine baselines spanning diffusion, latent diffusion, and flow-based paradigms: WDM [12], MD [18], 3DMD [31], MLDM [24], BS [29], MOTFM [32], and FlowLet [7]; the latter also provides age-conditioned variants of WDM and MOTFM, denoted WDMa and MOTFMa. All models are trained on the same generative training cohort. For evaluation, each model generated 3000 synthetic volumes; conditional models spanned the training age range linearly, unconditional samples were assigned ages from the training distribution.

Evaluation protocol. Quantitative global metrics in volumetric MRI can be skewed by the large proportion of empty background voxels, potentially hiding anatomical inconsistencies; therefore, we follow a three-level evaluation protocol proposed in [7]. (i) *Global distributional metrics*: FID and MMD quantify distributional alignment between generated and real MRIs using features extracted from a medical-pretrained ResNet-50 [3]. MS-SSIM is computed intra-set as the average pairwise similarity among generated samples, where lower values indicate greater sample diversity. (ii) *Brain Age Prediction (BAP)*: following [8], we train

Table 2: Generative quality. Best in **bold**, second underlined. [†]Unconditional model. All pairwise comparisons against WaveDiT-CFM 10 steps are statistically significant (Wilcoxon rank-sum, Bonferroni-corrected $p < 0.001$). Metrics are computed over 10 bootstrap resamples of 500 generated samples. Standard deviations ($\leq 10^{-3}$) are omitted for conciseness.

(a) External baselines					(b) WaveDiT (ours) & ablations						
<i>Baselines</i>	Method	Steps	FID ↓	MMD ↓	MS-SSIM ↓	<i>Ours</i>	Method	Steps	FID ↓	MMD ↓	MS-SSIM ↓
	3DMD [†]	1000	0.0307	0.00060	0.887		WaveDiT-RFM	10	0.1811	0.00719	0.862
	MD [†]	1000	0.0113	0.00041	0.898		WaveDiT-OTFM	10	0.0157	0.00056	0.830
	WDM [†]	1000	0.0045	0.00011	0.908		WaveDiT-CFM	10	0.0039	<u>0.00010</u>	<u>0.834</u>
	WDMa	1000	<u>0.0044</u>	0.00014	0.900	<i>Abl.</i>	Voxel-space	10	0.0045	0.00009	0.878
	MOTFM [†]	50	0.1154	0.00455	0.918		<i>w/o</i> Morpheus	10	0.0295	0.00053	0.862
	MOTFMa	50	0.1124	0.00433	0.848						
	MLDM	100	0.0704	0.00272	0.884						
	BS	–	0.0477	0.00178	0.852						
	FlowLet	10	0.0117	0.00040	0.869						

a separate 3D DenseNet for each generative method using its synthetic samples as data augmentation to predict chronological age from MRI. Each BAP model is then evaluated on the same held-out set of real subjects older than 44 years, and performance is reported as Mean Absolute Error in years. This age range is under-represented in OpenBHB, where data sparsity can affect prediction accuracy [9]. (iii) *Region-of-interest (ROI) analysis*: each synthetic volume and an age-matched real counterpart are independently segmented [13] into 95 cortical and subcortical regions. Region-wise intensity MAE, KL divergence, and Dice coefficient are computed over the union of real and synthetic segmentations and averaged across all ROIs, following the procedures of [7]. These protocols jointly ensure that evaluation captures age-specific anatomical fidelity that is robust to both distributional overfitting and diversity collapse. Unlike BAP, ROI metrics are independent of age balancing and directly assess anatomical fidelity.

Generative quality. Table 2 reports global distributional metrics, with visual examples in Figure 2. At 10 steps, WaveDiT-CFM achieves the lowest metrics among all methods, improving over the conditioned wavelet-based baseline FlowLet. Compared to WDM, which also operates in the wavelet domain but requires 1000 diffusion steps, WaveDiT-CFM achieves lower FID, indicating that the proposed flow-based formulation is well suited to high-resolution wavelet coefficients. Across internal objectives, CFM outperforms RFM [19] and OTFM [2], suggesting that conditional trajectories yield a more favorable balance between fidelity and efficiency than constant-velocity or optimal-transport in this setting.

Downstream evaluation. Table 3 summarizes BAP and ROI metrics. WaveDiT-CFM achieves the lowest BAP MAE, outperforming the reference model trained solely on real data under the same protocol and other conditional baselines when used for synthetic augmentation. It also achieves the best ROI-level scores, with

Table 3: Downstream evaluation. BAP: Test MAE in years ($\downarrow$); ROI: 95-region average. $\dagger$ Unconditional model. Best in **bold**, second underlined.

	Method	BAP $\downarrow$	iMAE $\downarrow$	KLD $\downarrow$	DICE $\uparrow$
	Real data	2.92 ± 4.08			
<i>Baselines</i>	3DMD †	3.10 ± 4.04	62.91 ± 17.07	2.97 ± 1.04	0.35 ± 0.15
	MD †	3.72 ± 3.24	59.33 ± 13.12	1.45 ± 0.55	0.05 ± 0.10
	WDM †	3.93 ± 6.17	52.65 ± 15.33	2.36 ± 1.30	0.36 ± 0.15
	WDMa	3.08 ± 4.62	48.57 ± 12.23	1.06 ± 0.45	0.33 ± 0.15
	MOTFM †	3.61 ± 5.59	61.00 ± 15.45	1.06 ± 0.58	0.42 ± 0.16
	MOTFMa	2.76 ± 3.50	50.68 ± 9.25	1.24 ± 0.63	0.41 ± 0.16
	MLDM	2.68 ± 3.76	48.72 ± 12.36	<u>0.97 ± 0.47</u>	0.32 ± 0.15
	BS	3.08 ± 3.95	48.84 ± 12.22	1.04 ± 0.46	0.33 ± 0.15
	FlowLet	2.72 ± 3.85	<u>45.18 ± 11.11</u>	1.18 ± 0.56	<u>0.44 ± 0.16</u>
<i>Ours</i>	WaveDiT-RFM	2.73 ± 4.13	62.96 ± 17.15	2.94 ± 1.06	0.33 ± 0.15
	WaveDiT-OTFM	<u>2.51 ± 3.23</u>	46.83 ± 11.81	1.02 ± 0.51	0.42 ± 0.16
	WaveDiT-CFM	2.44 ± 3.19	42.49 ± 12.96	0.96 ± 0.48	0.46 ± 0.16
<i>Abl.</i>	Voxel-space	3.04 ± 4.52	51.30 ± 12.74	1.05 ± 0.47	0.43 ± 0.16
	<i>w/o</i> Morpheus	2.98 ± 4.27	54.58 ± 13.77	1.01 ± 0.45	0.41 ± 0.16

lower iMAE and KL divergence and higher Dice than alternatives. Notably, WDM and MD achieve competitive FID while exhibiting reduced Dice and higher KLD, reinforcing that global metrics may underestimate discrepancies in regional brain structure and motivating the use of BAP and ROI-level analysis as complementary endpoints.

Ablations. The Morpheus module has a substantial impact on both global and downstream metrics. Removing Morpheus degrades FID and MMD and worsens BAP and ROI scores, indicating that uniform MSE is insufficient under the statistics of high-frequency wavelet bands, confirming the importance of state-aware precision control for stable training and improved anatomical fidelity. Among flow objectives, CFM provides the strongest overall trade-off: RFM exhibits degraded global and ROI performance, while OTFM improves MS-SSIM but does not match CFM on FID or anatomical scores, suggesting that CFM trajectories better align with wavelet-domain statistics under a limited step budget. Finally, the voxel-space variant, although similar in FID and MMD, yields higher BAP error and degraded ROI scores, indicating that matching global scores in voxel space does not guarantee preservation of clinically relevant anatomy.

4 Conclusion

We introduced WaveDiT, a wavelet-domain conditional flow matching model for 3D brain MRI synthesis that combines an HDiT backbone with the Morpheus state-aware uncertainty scheduler. In our experiments, WaveDiT-CFM achieved strong global distributional scores while also improving brain age prediction and ROI-level anatomical metrics compared to existing diffusion and flow-based baselines under a low-step sampling regime. Ablation studies suggest that both the

wavelet representation and state-aware uncertainty weighting contribute to stabilizing training and maintaining region-level anatomical plausibility. This work is limited to T1-weighted MRI and does not include expert reader studies, so we restrict our claims to quantitative proxies of anatomical fidelity and clinical utility. As future directions, we plan to explore WaveDiT on additional 3D imaging domains and modalities, such as CT, and to investigate richer conditioning schemes beyond age.

Acknowledgments. This work was partially supported by the following projects: “LIFE: the itaLian system wIde Frailty nEtnetwork”; We acknowledge the CINECA award under the ISCRA initiative (Projects IsCc1 SynBrain, IsCd1 FlowMRI, IsCd3 EMBRAIN) for the availability of high performance computing resources and support”; This work has been carried out while *Matteo Attimonelli* was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome in collaboration with *Politecnico Di Bari*.

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

References

1. Benkarim, O., Paquola, C., Park, B.y., Kebets, V., Hong, S.J., Vos de Wael, R., Zhang, S., Yeo, B.T., Eickenberg, M., Ge, T., et al.: Population heterogeneity in clinical cohorts affects the predictive accuracy of brain imaging. *PLoS biology* **20**(4), e3001627 (2022)
2. Chen, R.T.Q., Lipman, Y.: Flow matching on general geometries. In: *ICLR* (2024)
3. Chen, S., Ma, K., Zheng, Y.: Med3d: Transfer learning for 3d medical image analysis. *CoRR* **abs/1904.00625** (2019)
4. Chintapalli, S.S., Wang, R., Yang, Z., Tassopoulou, V., Yu, F., Bashyam, V., Erus, G., Chaudhari, P., Shou, H., Davatzikos, C.: Generative models of MRI-derived neuroimaging features and associated dataset of 18,000 samples. *Scientific Data* **11**(1), 1330 (Dec 2024)
5. Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Cnaan, M.W., Steves, C., Spector, T.D., Montana, G.: Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017)
6. 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*. OpenReview.net (2024)
7. Danese, D., et al.: Flowlet: Wavelet-based flow matching for efficient 3d brain mri synthesis. *arXiv preprint* (2025)
8. De Bonis, M.L.N., Fasano, G., Lombardi, A., Ardito, C., Ferrara, A., Di Sciascio, E., Di Noia, T.: Explainable brain age prediction: a comparative evaluation of morphometric and deep learning pipelines. *Brain Informatics* **11**(1), 33 (2024)
9. Dinsdale, N.K., Bluemke, E., Smith, S.M., Arya, Z., Vidaurre, D., Jenkinson, M., Namburete, A.I.: Learning patterns of the ageing brain in mri using deep convolutional networks. *NeuroImage* **224**, 117401 (2021)

10. Dufumier, B., Grigis, A., Victor, J., Ambroise, C., Frouin, V., Duchesnay, E.: Openbhb: a large-scale multi-site brain mri data-set for age prediction and debiasing. *NeuroImage* **263**, 119637 (2022). <https://doi.org/10.1016/j.neuroimage.2022.119637>, <https://baobablab.github.io/bhb/dataset>
11. Fonov, V., Evans, A., McKinstry, R., Almli, C., Collins, D.: Unbiased non-linear average age-appropriate brain templates from birth to adulthood. *NeuroImage* **47**, S102 (2009). [https://doi.org/10.1016/S1053-8119\(09\)70884-5](https://doi.org/10.1016/S1053-8119(09)70884-5), <https://www.sciencedirect.com/science/article/pii/S1053811909708845>, organization for Human Brain Mapping 2009 Annual Meeting
12. Friedrich, P., Wolleb, J., Bieder, F., Durrer, A., Cattin, P.C.: WDM: 3d wavelet diffusion models for high-resolution medical image synthesis. In: DGM4MICCAI@MICCAI. Springer (2024)
13. Henschel, L., Conjeti, S., Estrada, S., Diers, K., Fischl, B., Reuter, M.: FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* **219**, 117012 (2020)
14. Heo, B., Park, S., Han, D., Yun, S.: Rotary position embedding for vision transformer. In: ECCV (10). Springer (2024)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
16. Jenkinson, M., Bannister, P., Brady, M., Smith, S.: Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* **17**(2), 825–841 (2002). <https://doi.org/10.1006/nimg.2002.1132>
17. Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: NIPS. pp. 5574–5584 (2017)
18. Khader, F., Mueller-Franzes, G., Arasteh, S.T., Han, T., Haarbuerger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baeßler, B., Foersch, S., Stegmaier, J., Kuhl, C., Nebelung, S., Kather, J.N., Truhn, D.: Medical diffusion - denoising diffusion probabilistic models for 3d medical image generation. *CoRR* (2022)
19. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
20. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: ICLR. OpenReview.net (2023)
21. Marcus, D.S., Fotenos, A.F., Csernansky, J.G., Morris, J.C., Buckner, R.L.: Open access series of imaging studies: longitudinal MRI data in nondemented and demented older adults. *J. Cogn. Neurosci.* (2010), sites.wustl.edu/oasisbrains/
22. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarbuerger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., Kather, J.N., Truhn, D.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports* **13**(1), 12098 (Jul 2023)
23. Petersen, R.C., Aisen, P.S., Beckett, L.A., Donohue, M.C., Gamst, A.C., Harvey, D.J., Jack, Jr, C.R., Jagust, W.J., Shaw, L.M., Toga, A.W., Trojanowski, J.Q., Weiner, M.W.: Alzheimer’s disease neuroimaging initiative (ADNI): clinical characterization. *Neurology* **74**(3), 201–209 (Jan 2010), <https://adni.loni.usc.edu/>
24. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models (2022)
25. Rahman, M.T., Orka, N.A., Khan, A., Liò, P., Moni, M.A.: Understanding neurocognition with deep learning and mri: A systematic review. *IEEE Transactions on Cognitive and Developmental Systems* (2025)

26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. IEEE (2022)
27. Seitzer, M., Tavakoli, A., Antic, D., Martius, G.: On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In: ICLR (2022)
28. Smith, S.M.: Fast robust automated brain extraction. *Hum. Brain Mapp.* **17**(3), 143–155 (Nov 2002)
29. Tudosi, P., Pinaya, W.H.L., Costa, P.F.D., Dafflon, J., Patel, A., Borges, P., Fernandez, V., Graham, M.S., Gray, R.J., Nachev, P., Ourselin, S., Cardoso, M.J.: Realistic morphology-preserving generative modelling of the brain. *Nat. Mac. Intell.* **6**(7), 811–819 (2024)
30. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* (2010)
31. Wang, H., Liu, Z., Sun, K., Wang, X., Shen, D., Cui, Z.: 3d meddiffusion: A 3d medical latent diffusion model for controllable and high-quality medical image generation. *IEEE Trans. Medical Imaging* **44**(12), 4960–4972 (2025)
32. Yazdani, M., Medghalchi, Y., Ashrafian, P., Hacihaliloglu, I., Shahriari, D.: Flow matching for medical image synthesis: Bridging the gap between speed and quality. *arXiv preprint arXiv:2503.00266* (2025)
33. Zhang, B., Sennrich, R.: Root mean square layer normalization. In: NeurIPS. pp. 12360–12371 (2019)
34. Zhang, X., Pak, D.H., Ahn, S.S., Li, X., You, C., Staib, L.H., Sinusas, A.J., Wong, A.L.N., Duncan, J.S.: Heteroscedastic uncertainty estimation framework for unsupervised registration. In: MICCAI (2). Springer (2024)