WFM: 3D Wavelet Flow Matching for Ultra-Fast Multi-Modal MRI Synthesis

Yalcin Tur[1] YALCINTR@STANFORD.EDU
Mihajlo Stojkovic[1] MSTOJKOV@STANFORD.EDU
Ulas Bagci[2] ULAS.BAGCI@NORTHWESTERN.EDU
[1] *Department of Computer Science, Stanford University, Stanford, CA, USA.*
[2] *Machine and Hybrid Intelligence Lab, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA.*

Editors: Under Review for MIDL 2026

Abstract

Diffusion models have achieved remarkable quality in multi-modal MRI synthesis, but their computational cost—hundreds of sampling steps and separate models per modality—limits clinical deployment. We observe that this inefficiency stems from an unnecessary starting point: diffusion begins from pure noise, discarding the structural information already present in available MRI sequences. We propose WFM (Wavelet Flow Matching), which instead learns a direct flow from an *informed prior*—the mean of conditioning modalities in wavelet space—to the target distribution. Because the source and target share underlying anatomy and differ primarily in contrast, this formulation enables accurate synthesis in just 1-2 integration steps. A single 82M-parameter model with class conditioning synthesizes all four BraTS modalities (T1, T1c, T2, FLAIR), replacing four separate diffusion models totaling 326M parameters. On BraTS 2024, WFM achieves 26.8 dB PSNR and 0.94 SSIM—within 1-2 dB of diffusion baselines—while running 250-1000x faster (0.16-0.64s vs. 160s per volume). This speed-quality trade-off makes real-time MRI synthesis practical for clinical workflows. Code will be released upon acceptance of the paper.

Keywords: MRI synthesis, flow matching, wavelet transform, diffusion models, multi-modal imaging

1. Introduction

Accurate brain tumor analysis requires multiple MRI contrasts—T1-weighted, T1 with contrast enhancement (T1c), T2-weighted, and T2-FLAIR—each revealing distinct tissue properties essential for segmentation and treatment planning (Baid et al., 2021). In practice, however, complete acquisitions are frequently unavailable: scan time constraints, patient motion, or imaging artifacts often leave one or more modalities missing. Synthesizing these missing sequences from available ones has therefore become an active research focus, with recent diffusion-based methods achieving impressive fidelity. Yet a fundamental barrier remains: these methods are too slow for clinical use.

Diffusion models have set the quality standard for medical image synthesis (Friedrich et al., 2024a; Kim and Park, 2024; Özbey et al., 2023). The conditional Wavelet Diffusion Model (cWDM) demonstrated that operating in wavelet space enables memory-efficient 3D synthesis at full resolution (Friedrich et al., 2024a). However, these methods inherit

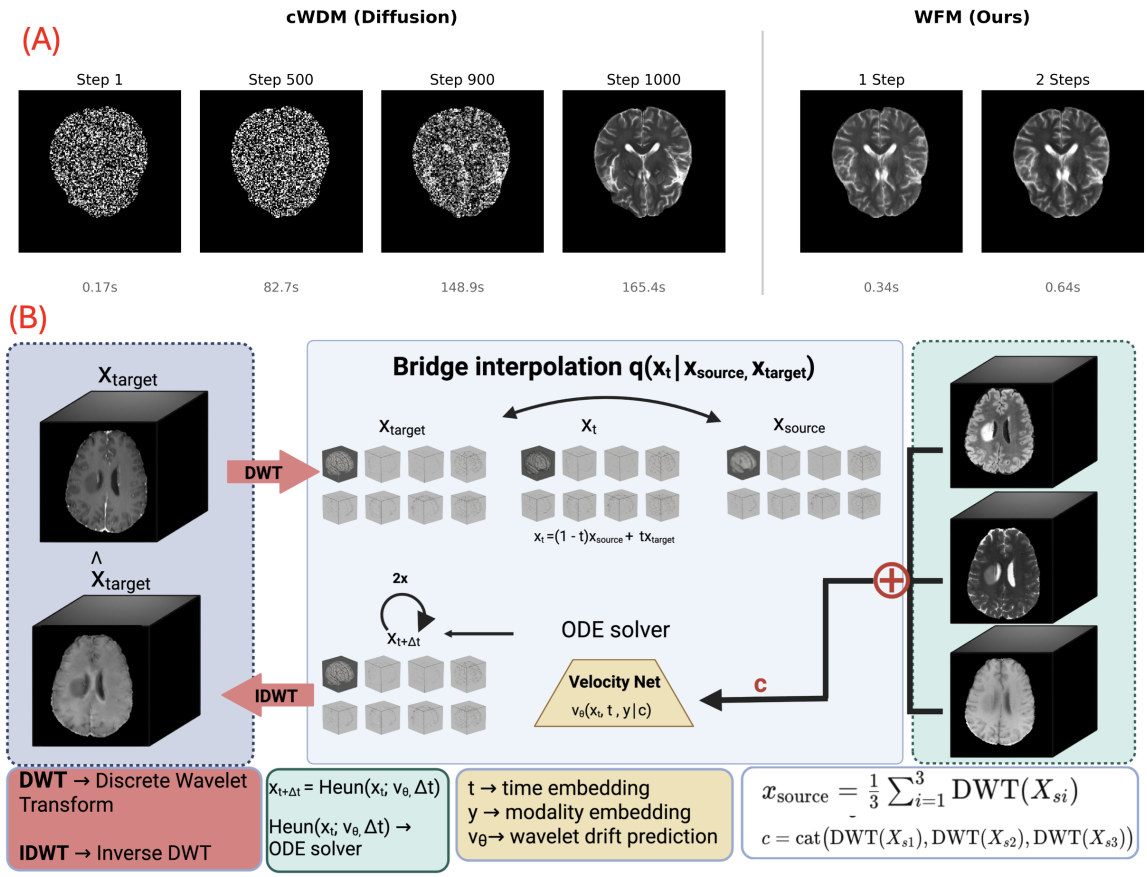

Figure 1: **(A)** Overview of WFM. Three conditioning modalities are transformed to wavelet space and averaged to form the informed source $\mathbf{x}_{\text{source}}$; their concatenation forms the condition $\mathbf{c}$. The velocity network predicts the drift from source to target, conditioned on timestep t and target modality y. An ODE solver (Euler or Heun) integrates the velocity field in 1-2 steps. The inverse wavelet transform (IDWT) produces the final synthesized volume $\hat{\mathbf{X}}_{\text{target}}$. **(B)** Comparison of synthesis speed. cWDM requires 1000 denoising steps (165.4s) to transform noise into a valid MRI. WFM produces comparable quality in 1-2 steps (0.34-0.64s) by starting from an informed prior rather than noise.

a fundamental inefficiency from the diffusion framework: they begin from pure Gaussian noise. Reconstructing meaningful anatomical structure from noise requires iterative refinement—typically 1000 denoising steps, translating to approximately 160 seconds per volume. Moreover, because each target modality defines a distinct generation task, existing approaches train separate models for each contrast, resulting in four independent networks totaling 326M parameters for the BraTS protocol. Together, these costs place diffusion-based synthesis outside the realm of real-time clinical deployment.

Our key observation is that the diffusion starting point is unnecessarily uninformative. In multi-modal MRI synthesis, source and target modalities capture the same anatomy with different contrast weightings—T1 and T2 images of the same brain differ in intensity

mapping, not in structure. The mean of available modalities, therefore, provides a strong prior: it already contains ventricle boundaries, cortical folding, and lesion locations. Rather than reconstructing this structure from noise, we can learn a direct transformation from the **informed prior** to the target contrast. This insight aligns with recent advances in flow matching (Lipman et al., 2023) and bridge-based diffusion (Liu et al., 2023; Li et al., 2023), which demonstrate that informative starting points enable few-step generation.

We propose **WFM (Wavelet Flow Matching)**, a method that combines these principles with 3D wavelet-domain processing for efficient multi-modal MRI synthesis. Our contributions are threefold:

1. **Informed-prior flow matching.** We formulate synthesis as learning a velocity field from the mean of conditioning modalities (in wavelet space) to the target distribution. Because the prior already encodes anatomical structure, the model learns only the contrast transformation, enabling accurate synthesis in 1-2 integration steps rather than 1000.

2. **Unified multi-modality architecture.** A single 82M-parameter network with class conditioning synthesizes all four BraTS modalities, replacing four separate models (326M total) while sharing learned anatomical representations across tasks.

3. **Clinically viable speed.** WFM synthesizes a full 240×240×155 volume in 0.16-0.64 seconds—a 250-1000x speedup over cWDM (**Figure** 1A)—bringing MRI synthesis into the realm of interactive clinical workflows.

2. Related Work

2.1. Diffusion Models for Medical Image Synthesis Diffusion models have become the dominant paradigm for high-fidelity medical image synthesis, displacing GANs where quality is paramount. Friedrich et al. (2024b) addressed the memory challenge of 3D synthesis by operating in wavelet space, and their conditional extension cWDM (Friedrich et al., 2024a) demonstrated state-of-the-art quality on BraTS. Kim and Park (2024) combined SPADE normalization with latent diffusion; Özbey et al. (2023) proposed SynDiff for unsupervised translation. Despite architectural innovations, all these methods require hundreds to thousands of sampling steps. Acceleration techniques—DDIM (Song et al., 2020), progressive distillation (Salimans and Ho, 2022), consistency models (Song et al., 2023)—reduce step counts but retain noise as the starting point. WFM instead changes the starting point itself: by beginning from an informed prior, single-step integration becomes accurate without distillation.

2.2. Flow Matching and Bridge Methods Flow matching (Lipman et al., 2023) learns continuous-time flows between distributions, offering simpler training objectives than diffusion. Several works exploit informative starting distributions: I^2SB (Image-to-Image Schrödinger Bridge) Liu et al. (2023) formulates translation as a Schrödinger bridge, BBDM Li et al. (2023) uses Brownian bridge processes, and DSBM Shi et al. (2023) provides theoretical foundations. WFM makes a simplifying assumption suited to multi-modal MRI: because source and target share anatomy and differ only in contrast, a linear interpolation path suffices, yielding single-step Euler integration without complex bridge formulations.

2.3. GAN-based Medical Image Translation Before diffusion, GANs dominated medical synthesis—Pix2Pix (Isola et al., 2017), CycleGAN (Zhu et al., 2017), ResViT Dalmaz et al. (2022)—offering fast inference but suffering from training instability in 3D. WFM achieves GAN-like speed through flow matching while avoiding adversarial training. For memory efficiency, we inherit the wavelet-domain formulation of (Friedrich et al., 2024b): the 3D Haar transform reduces spatial dimensions while preserving information losslessly, enabling full-resolution synthesis. Finally, we adopt a unified multi-task architecture with class conditioning, reducing parameters from 326M (four cWDM models) to 82M while sharing anatomical representations (Chartsias et al., 2018).

3. Methods

Problem Formulation. We consider the task of synthesizing a missing MRI modality from a set of available sequences. Let $\{\mathbf{X}_{s_1}, \mathbf{X}_{s_2}, \mathbf{X}_{s_3}\} \in \mathbb{R}^{D\times H\times W}$ denote the three out of four BraTS modalities (T1, T1c, T2, FLAIR), where each $X^{(i)} \in \mathbb{R}^{DHW}$ is a 3D volume with depth D, height H, and width W. Given three available modalities $\{\mathbf{X}_{s_1}, \mathbf{X}_{s_2}, \mathbf{X}_{s_3}\}$ as conditioning inputs, our goal is to synthesize the missing target X_t. Prior diffusion-based approaches train separate models for each target modality, yielding four independent networks. We instead learn a single unified model f that synthesizes any of the four modalities, using a class label $y \in \{0, 1, 2, 3\}$ to specify which modality to generate. However, this formulation assumes exactly three conditioning modalities are available—a constraint we discuss in Section 5. The key departure from diffusion-based synthesis is the choice of starting distribution. Rather than generating X_t by denoising from Gaussian noise, we construct an informed prior from the conditioning modalities and learn a direct transformation to the target. The following sections formalize this approach.

3.1. Flow Matching with Informed Prior

Background. Standard diffusion models define a forward process that progressively corrupts data toward isotropic Gaussian noise, then learn to reverse this process. Generation proceeds by sampling from $\mathcal{N}(0, \mathbf{I})$ and iteratively denoising. The limitation is fundamental: pure noise contains no information about the target, so the model must reconstruct all structure from scratch—a task requiring many refinement steps. Flow matching (Lipman et al., 2023) offers an alternative formulation. Rather than learning a denoising process, flow matching learns a velocity field $\mathbf{v}(x, t)$ that transports samples from a source distribution to a target distribution along continuous paths. The framework is flexible: any differentiable path connecting the source and target defines a valid flow.

Informed source construction. Our key insight is that the source distribution need not be uninformative noise. In multi-modal MRI synthesis, the conditioning modalities $\{\mathbf{X}_{s_1}, \mathbf{X}_{s_2}, \mathbf{X}_{s_3}\}$ already encode the patient's anatomy—the same ventricles, cortical folds, and lesions that will appear in the target X_t, differing only in intensity mapping. We construct an informed source by averaging the conditioning modalities in wavelet space:

$$\mathbf{x}_{\text{source}} = \frac{1}{3}\sum_{i=1}^{3} \text{DWT}(\mathbf{X}_{s_i}), \tag{1}$$

where DWT denotes the 3D discrete wavelet transform (detailed in Section 3.4). The mean operation is motivated by two considerations. First, averaging suppresses modality-specific intensity variations while preserving shared anatomical structure—edges, boundaries, and spatial organization that are consistent across sequences. Second, the mean provides a simple, parameter-free aggregation that introduces no additional learned components. Alternative aggregation strategies—learned attention, concatenation with reduction, or modality-specific weighting—could potentially improve performance but would increase model complexity. We find that simple averaging suffices for the BraTS synthesis task (see **Figure 1** for the overall pipeline).

Linear interpolation path. Given source $\mathbf{x}_{\text{source}}$ and target $\mathbf{x}_{\text{target}} = DWT(\mathbf{X_t})$, we define a linear interpolation: $\mathbf{x}_t = (1-t)\mathbf{x}_{\text{source}} + t\mathbf{x}_{\text{target}}$, where $t \in [0, 1]$, This path has a constant velocity:

$$\mathbf{v} = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_{\text{target}} - \mathbf{x}_{\text{source}}. \tag{2}$$

The linear path is the simplest choice and corresponds to optimal transport with quadratic cost when source and target are point masses. For MRI synthesis, where source and target are structurally aligned, this direct path is appropriate: there is no need for the curved trajectories that optimal transport would prescribe for distant, misaligned distributions. In practice, the learned velocity $\mathbf{v}$ is an approximation, but the structural alignment between source and target ensures that errors are small—enabling accurate synthesis in 1-2 steps rather than hundreds.

3.2. Training Objective

We train a neural network f_θ to predict the constant velocity $\mathbf{v} = \mathbf{x}_{\text{target}} - \mathbf{x}_{\text{source}}$.The training objective is a simple regression loss:

$$\mathcal{L} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\| f_\theta(\tilde{\mathbf{x}}_t, \mathbf{c}, t, y) - (\mathbf{x}_{\text{target}} - \mathbf{x}_{\text{source}}) \|^2 \right] \tag{3}$$

where $t \sim \mathcal{U}(0, 1)$ is a uniformly sampled timestep, $y \in \{0, 1, 2, 3\}$ is the target modality class, and $\mathbf{c}$ is the conditioning information (defined below).

Conditioning representation. The condition $\mathbf{c}$ consists of the concatenated wavelet coefficients of all three source modalities. This provides the model with full access to each conditioning modality separately, enabling it to learn modality-specific contrast relationships rather than relying solely on the averaged source.

Stochastic regularization. During training, we add noise to the interpolant:

$$\tilde{\mathbf{x}}_t = \mathbf{x}_t + \sigma\sqrt{t(1-t)}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

The noise schedule σ has a specific motivation. At the endpoints ($t = 0$ and $t = 1$), the noise vanishes, ensuring that the model sees clean source and target samples. At intermediate timesteps, noise is maximal at $t = 0.5$, where the interpolant is equidistant from both endpoints and perturbations are least disruptive to the learning signal. This regularization serves two purposes. First, it prevents the model from memorizing the deterministic interpolation path, encouraging generalization to unseen source-target pairs. Second, it provides robustness during inference: small errors in the predicted velocity do not compound catastrophically because the model has been trained on perturbed trajectories. We set $\sigma = 0.5$

based on preliminary experiments. Higher values degrade quality by obscuring the target velocity signal; lower values reduce regularization benefit.

3.3. Unified Multi-Modality Architecture

A naive approach to multi-modal synthesis trains separate models for each target modality-four networks for the BraTS protocol, totaling 326M parameters (81.5M x 4). This is inefficient: the networks learn redundant anatomical representations, and knowledge cannot transfer across synthesis tasks. We instead train a single unified model f_θ (3D U-Net with class conditioning) that synthesizes any target modality, specified by a class label y. This reduces total parameters to 81.5M (a 4x reduction) while enabling the network to share learned representations across all synthesis directions. The class label y is embedded and added to the timestep embedding: $\mathbf{e} = \text{TimeEmbed}(t) + \text{ClassEmbed}(y)$. This vector modulates the network through adaptive normalization: each residual block scales and shifts its normalized activations based on $\mathbf{e}$, enabling the same weights to produce different behaviors for different target modalities and timesteps. The model takes as input the concatenation of the noisy interpolant $\tilde{\mathbf{x}}_t$ (8 channels) with all three conditioning modalities in wavelet space (24 channels), totaling 32 input channels. 3D U-Net with base channels 64, channel multipliers (1, 2, 2, 4, 4), 2 residual blocks per level. Total parameters: 81.5M.

At inference time, we synthesize a target modality by integrating the learned velocity field from source to target. The number of function evaluations (NFE)—forward passes through the network—determines both quality and speed. We consider two ODE solvers: Euler (first-order) and Heun (second-order Runge-Kutta). For single-step Euler (NFE=1), this directly yields the target $\hat{\mathbf{x}}_{\text{target}} = \mathbf{x}_{\text{source}} + f_\theta(\mathbf{x}_{\text{source}}, \mathbf{c}, t = 0, y)$. For higher accuracy, Heun's method uses a predictor-corrector scheme (2 forward calls per step):

$$\mathbf{v}_1 = f_\theta(\mathbf{x}_t, \mathbf{c}, t, y), \tag{5}$$

$$\tilde{\mathbf{x}}_{t+\Delta t} = \mathbf{x}_t + \mathbf{v}_1 \cdot \Delta t, \tag{6}$$

$$\mathbf{v}_2 = f_\theta(\tilde{\mathbf{x}}_{t+\Delta t}, \mathbf{c}, t + \Delta t, y), \tag{7}$$

$$\mathbf{x}_{t+\Delta t} = \mathbf{x}_t + \frac{\mathbf{v}_1 + \mathbf{v}_2}{2} \cdot \Delta t. \tag{8}$$

The final output is obtained by applying the inverse wavelet transform.

4. Experiments

Dataset and preprocessing: We evaluate on BraTS 2024 (Baid et al., 2021), a multi-institutional dataset of brain MRI volumes from glioma patients. Each case includes four co-registered modalities: T1-weighted (T1), T1 with gadolinium contrast enhancement (T1c), T2-weighted (T2), and T2-FLAIR. Volumes are resampled to 1mm^3 isotropic resolution with dimensions 240×240×155 voxels. We use the official BraTS 2024 split: 1,032 volumes for training and 219 for validation. The validation set serves as our test set since BraTS withholds ground truth for the official test partition. All reported metrics are computed on this 219-volume validation set. Each modality is independently normalized to zero mean and unit variance within the brain mask. We do not apply additional intensity standardization

(e.g., histogram matching) to preserve the natural contrast variations that the synthesis model must learn to handle.

Implementation details and baselines: Training uses AdamW optimizer with learning rate 10^{-5}, batch size 4, for 50K iterations on a single NVIDIA A100. During training, the target modality is randomly selected each step. We use $\sigma = 0.5$ for regularization noise. Our primary comparison is against cWDM (Friedrich et al., 2024a), the current state-of-the-art for 3D MRI synthesis on BraTS. cWDM operates in wavelet space like WFM but uses a diffusion formulation requiring 1000 sampling steps. We use the authors' released checkpoints, which consist of four separate models (one per target modality) totaling 326M parameters. We acknowledge that our baseline comparison is limited to a single method. Comprehensive comparison against GAN-based approaches (Pix2Pix-3D, ResViT) and other flow-based methods (I^2SB, LBM) would strengthen the evaluation but requires substantial additional implementation effort for 3D medical volumes, which we leave to future work.

Metrics and their potential limitations: We report two complementary metrics computed within the brain mask: PSNR and SSIM. PSNR measures pixel-wise reconstruction accuracy in decibels. Higher is better. SSIM measures perceptual similarity based on luminance, contrast, and structure. Range [0, 1]; higher is better. Both metrics are computed per-volume and averaged across the validation set. PSNR and SSIM assess pixel-level fidelity but do not directly measure clinical utility. A more complete evaluation would include: (1) perceptual metrics like LPIPS or FID adapted for 3D medical images, (2) downstream task performance (e.g., tumor segmentation accuracy using synthetic vs. real modalities), and (3) radiologist evaluation of diagnostic quality. We discuss these limitations in Section 5.5.

4.1. Main Results

Table 1 presents the quantitative comparison between WFM and cWDM across all four BraTS modalities. Overall, WFM achieves 26.8 dB average PSNR and 0.94 SSIM with 1-2 function evaluations, compared to cWDM's 28.4 dB PSNR and 0.95 SSIM with 1000 evaluations. The quality gap of 1.6 dB PSNR represents a meaningful but bounded degradation: in perceptual terms, differences below 1 dB are typically imperceptible, while differences above 3 dB become visually obvious. The 1.6 dB gap falls in the "noticeable upon close inspection" range. All ablation studies are presented in the appendix due to space constraints.

Per-modality analysis. Performance varies substantially across target modalities. T1c synthesis shows the smallest gap (0.78 dB), likely because contrast enhancement patterns in T1c are partially predictable from the non-contrast T1 signal combined with T2/FLAIR tumor delineation. T1 synthesis shows the largest gap (2.12 dB), suggesting that reconstructing T1's specific intensity relationships from T1c/T2/FLAIR is more challenging than the reverse direction.

Integration method comparison. Heun's method with 2 steps (NFE=4) provides marginal improvement over single-step Euler (NFE=1): 26.80 vs. 26.72 dB average PSNR. This confirms that the learned velocity field is approximately constant—additional integration steps yield diminishing returns because the true trajectory is nearly linear.

Unified vs. separate models. The single unified WFM model matches the quality achievable with four hypothetically separate WFM models (we verified this in preliminary experiments, finding *leq* 0.1 dB difference). This validates that class conditioning effectively specializes the shared architecture for each synthesis direction, with the added benefit of cross-task representation sharing.

Statistical considerations. Validation-set standard deviations are substantial (±2-3 dB) due to patient variability; the WFM-cWDM gap is consistent and significant.

Table 1: Comparison with cWDM on BraTS 2024 validation set. WFM achieves competitive quality with 4× fewer parameters and 250-1000x faster inference.

Method	Models	Params	NFE	T1	T1c	T2	FLAIR	Time
PSNR (dB) ↑								
cWDM	4	326M	1000	29.74	27.31	28.86	27.83	160s
WFM (Euler, 1)	1	82M	1	27.41	26.40	27.15	25.91	0.16s
WFM (Heun, 1)	1	82M	1	27.51	26.39	27.24	25.91	0.32s
WFM (Heun, 2)	1	82M	2	27.62	26.53	26.90	26.13	0.64s
SSIM ↑								
cWDM	4	326M	1000	0.956	0.936	0.952	0.935	160s
WFM (Euler, 1)	1	82M	1	0.947	0.930	0.943	0.928	0.16s
WFM (Heun, 1)	1	82M	1	0.948	0.930	0.944	0.928	0.32s
WFM (Heun, 2)	1	82M	2	0.948	0.932	0.944	0.930	0.64s

4.2. Efficiency Analysis

Table 2 summarizes the computational efficiency of WFM compared to cWDM. WFM achieves 250-1000x speedup depending on the integration scheme. The speedup is multiplicative: each cWDM denoising step requires one forward pass through an 81.5M parameter network, and 1000 steps accumulate to 160 seconds per volume. WFM's informed prior reduces this to 1-4 forward passes.

Parameter efficiency. cWDM requires four separate models (one per target modality), totaling 326M parameters. WFM uses a single unified model with 81.5M parameters—a 4x reduction. This impacts storage (one checkpoint vs. four), deployment complexity (one model to serve), and potentially training cost (though we did not compare training times directly). Both methods operate in wavelet space with similar per-forward-pass memory requirements 12GB for batch size 1 on A100). The memory advantage of WFM is modest; the primary benefit is wall-clock time.

4.3. Qualitative Results

Figure 2 (and 3 in the Appendix D) present representative synthesis results across multiple test cases. We analyze both successful syntheses and failure modes.

Table 2: Efficiency comparison. WFM achieves 4× parameter reduction and 250-1000x speedup. Each forward pass takes ∼160ms on an NVIDIA A100.

Method	Models	Params	Fwd Calls	Time	Speedup
cWDM (Friedrich et al., 2024a)	4 separate	326M	1000	160s	1×
WFM (Euler, 1)	1 unified	82M	1	0.16s	1000×
WFM (Heun, 1)	1 unified	82M	2	0.32s	500×
WFM (Heun, 2)	1 unified	82M	4	0.64s	250×

Anatomical fidelity and successful tumor synthesis. WFM consistently preserves major anatomical structures across all modalities. Ventricular boundaries, cortical folding patterns, and white-gray matter interfaces align closely with ground truth. This confirms that the informed prior successfully transfers structural information, and the learned velocity field accurately transforms contrast without distorting anatomy. Figure 2(right) (Samples 46 and 72) demonstrates that WFM can faithfully synthesize challenging pathological regions. Tumor boundaries, peritumoral edema, and heterogeneous enhancement patterns are preserved when sufficient structural cues exist in the conditioning modalities. The T2/FLAIR hyperintensity that delineates edema transfers effectively to synthesized T1c, enabling plausible enhancement prediction.

Failure modes. Figure 2(left) (Sample 30) illustrates under-detailed tumor cores where T1c enhancement depends on blood-brain barrier disruption—not observable in non-contrast sequences. This is fundamental to the task, not WFM-specific, though diffusion's stochastic sampling may better capture such uncertainty.

5. Discussion and Concluding Remarks

We presented WFM, a method that reframes multi-modal MRI synthesis as flow matching from an informed prior rather than denoising from noise. Because source and target modalities share underlying anatomy, the mean of conditioning modalities provides a starting point requiring only contrast transformation. The result is a 250-1000x speedup (0.16-0.64s vs. 160s per volume) with a single 82M-parameter model replacing four separate diffusion networks, at a cost of 1.6 dB PSNR (26.8 vs. 28.4 dB). Future work includes extending to cross-modality pairs with weaker structural alignment (CT-to-MRI), downstream validation on tumor segmentation, and uncertainty quantification for safe clinical deployment. More broadly, WFM illustrates that task-specific structure can dramatically reduce computational requirements—the assumption that generation must begin from noise is a modeling choice, not a necessity.

Why Informed Priors Enable Few-Step Synthesis. The speedup (from 1000 steps to 1-2) stems from initialization: conditioning modalities already encode anatomy, so the mean provides a starting point structurally close to the target. The model learns contrast transformation rather than structural reconstruction—explaining why NFE>2 provides diminishing returns (Table 3).

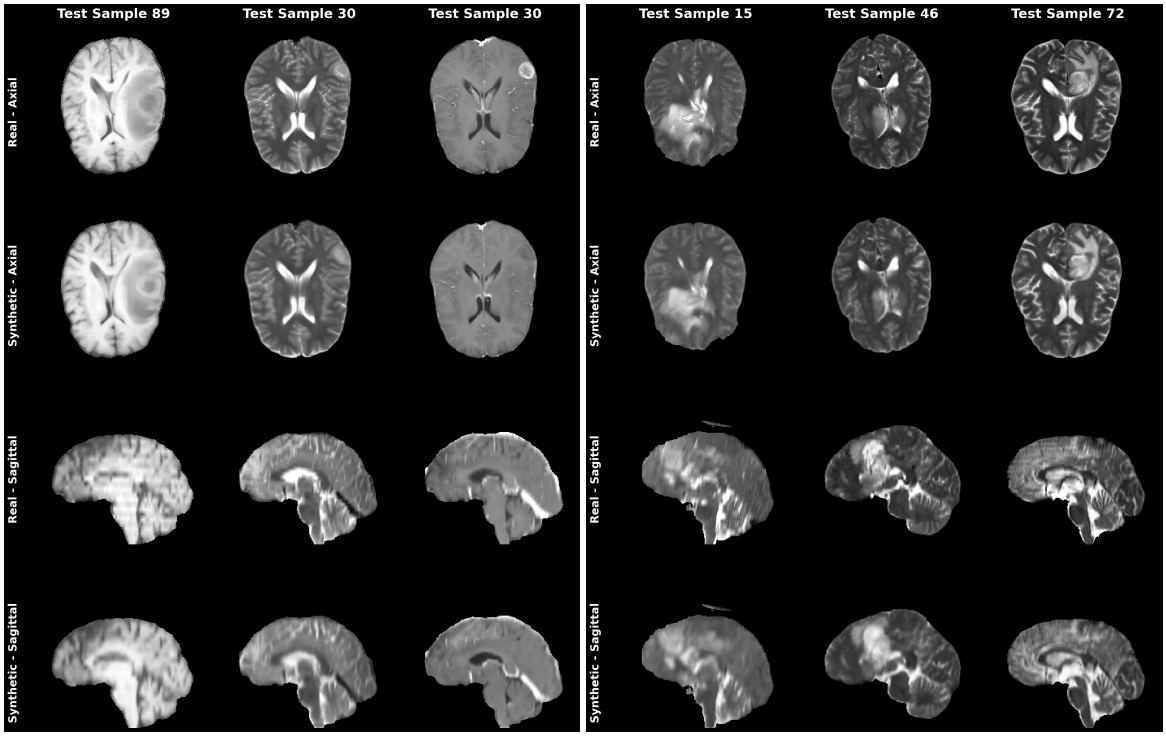

Figure 2: **Left:** Qualitative results including a failure case. Sample 89 (left) shows successful synthesis. Sample 30 (middle, right) illustrates a limitation: the model captures overall anatomy but produces less detailed tumor regions compared to ground truth, likely due to the unpredictable nature of contrast enhancement patterns. **Right:** Successful tumor synthesis cases. Samples 46 and 72 demonstrate accurate preservation of tumor boundaries and enhancement patterns. The model captures both the anatomical context and pathological features, showing that WFM can faithfully synthesize challenging tumor regions when sufficient structural information is present in the conditioning modalities.

Quality-speed trade-off. WFM achieves 26.8 dB PSNR versus cWDM's 28.4 dB—a 1.6 dB gap. Whether this is acceptable depends on application: for real-time clinical review, sub-second synthesis with 0.94 SSIM may be preferable to 160-second waits; for maximum-fidelity research, the gap matters more. The gap varies by modality (T1c: 0.78 dB; T1: 2.12 dB), reflecting varying difficulty of contrast prediction.

Limitations. Our evaluation has several limitations: (1) Single dataset—BraTS 2024 glioma patients; generalization to other pathologies is unvalidated. (2) No downstream evaluation—we do not test whether synthetic modalities improve segmentation. (3) Fixed conditioning—WFM assumes exactly three available modalities. (4) Single baseline-comparison with GANs and other flow methods would strengthen claims. (5) Failure modes—WFM struggles with unpredictable enhancement patterns in tumor cores, where T1c contrast depends on blood-brain barrier disruption not observable in non-contrast sequences. Extended discussion of alternative formulations, wavelet vs. latent processing, and clinical deployment considerations appears in the Appendix.

Acknowledgments

This study is partially supported by the following NIH grants: R01HL171376 and U01CA268808.

References

Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 29086–29098, October 2025.

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37(3):803–814, 2018. doi: 10.1109/TMI.2017.2764326.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10): 2598–2614, 2022.

Paul Friedrich, Alicia Durrer, Julia Wolleb, and Philippe C. Cattin. cwdm: Conditional wavelet diffusion models for cross-modality 3d medical image synthesis. *arXiv preprint arXiv:2411.17203*, 2024a.

Paul Friedrich, Julia Wolleb, Florentin Bieder, Alicia Durrer, and Philippe C Cattin. Wdm: 3d wavelet diffusion models for high-resolution medical image synthesis. In *MICCAI Workshop on Deep Generative Models*, pages 11–21. Springer, 2024b.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter conference on applications of computer Vision*, pages 7604–7613, 2024.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1952–1961, June 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling, 2023. URL https://arxiv.org/abs/2210.02747.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I ̋2 sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

Hao Phung, Quan Dao, and Anh Tran. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10199–10208, June 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models, 2022. URL https://arxiv.org/abs/2202.00512.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36: 62183–62223, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

Appendix A. Ablation Studies

We conduct ablation experiments to validate key design choices. Unless otherwise specified, ablations use Heun integration with 1 step (NFE=2). Table 3 shows that NFE=1-2 achieves optimal quality. Additional steps provide no benefit, consistent with our formulation where the model learns an approximately constant velocity field.

Table 3: Effect of number of function evaluations (NFE). NFE=2 with Heun's method achieves optimal quality.

Solver	NFE	T1	T1c	T2	FLAIR	Avg PSNR
Euler	1	27.41	26.40	27.15	25.91	26.72
Heun	1	27.51	26.39	27.24	25.91	26.76
Heun	2	27.62	26.53	26.90	26.13	26.80

Appendix B. Wavelet-Domain Processing

Operating directly on full-resolution 3D volumes (240×240×155 voxels at 32-bit precision ≈ 34 MB per volume) is memory-prohibitive for deep networks with batch processing and intermediate activations. Following Friedrich et al. (2024b,a); Phung et al. (2023), we

process volumes in wavelet space, where spatial dimensions are reduced while information is preserved.

3D Discrete Wavelet Transform. The 3D Haar wavelet transform decomposes a volume $\mathbf{X} \in \mathbb{R}^{D \times H \times W}$ into 8 subbands:

$$\text{DWT}(\mathbf{X}) \in \mathbb{R}^{8 \times \frac{D}{2} \times \frac{H}{2} \times \frac{W}{2}} \tag{9}$$

reducing spatial dimensions by $2\times$ in each axis while preserving multi-scale information through the LLL (low-frequency) and high-frequency subbands. Each subband captures different frequency content: LLL: Low-frequency approximation (coarse structure, smooth regions), LLH, LHL, HLL: Mixed frequency (edges along one axis), LHH, HLH, HHL: Higher frequency (edges along two axes), HHH: Highest frequency (corners, fine texture). The transform is orthogonal and perfectly invertible: $IDWT(DWT(\mathbf{X})) = \mathbf{X}$ with no information loss. This distinguishes wavelet processing from learned latent spaces, which introduce reconstruction error.

Choice of Haar and Processing pipeline. We use Haar wavelets for their computational simplicity: the transform requires only additions and subtractions, with no floating-point multiplications. More sophisticated wavelet families (Daubechies, biorthogonal) provide better frequency localization but increase computational cost without clear benefit for our synthesis task, where the learned network can compensate for transform limitations. All operations—source construction, interpolation, velocity prediction, and ODE integration—occur in wavelet space. Only the final output is transformed back to image space via IDWT for evaluation and visualization.

Appendix C. Implementation Details

C.1. Network Architecture

- Input channels: 32 (8 wavelet + 24 condition wavelet)
- Output channels: 8 (wavelet subbands)
- Base channels: 64
- Channel multipliers: (1, 2, 2, 4, 4)
- Residual blocks per level: 2
- Normalization: GroupNorm with 32 groups
- Total parameters: 81,512,072

C.2. Training Protocol

At each training step, we randomly select one of the four modalities as the target y, with the remaining three serving as conditions. This ensures balanced exposure to all synthesis directions. The model receives no explicit information about which specific modalities are conditioning inputs—only their wavelet representations and the target class label.

- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)

- Learning rate: 10^{-5}
- Batch size: 4
- Iterations: 50,000
- Gradient clipping: max norm 1.0
- Timestep sampling: $t \sim \mathcal{U}(0, 1)$
- Regularization noise: $\sigma = 0.5$

C.3. Inference Protocol

For Heun's method with NFE=2:

- Timesteps: $t \in \{0, 0.5, 1.0\}$
- Step size: $\Delta t = 0.5$

Timing (NVIDIA A100):

- Single forward pass: 160ms
- Euler NFE=1: 160 ± 0.05ms (1 call)
- Heun NFE=1: 321 ± 2.0ms (2 calls)
- Heun NFE=2: 640 ± 0.3ms (4 calls)
- cWDM (1000 steps): 160s

Appendix D. Additional Qualitative Results

Figure 3 shows additional synthesis results across three more test samples.

Appendix E. Extended Discussion

Regional error analysis. Across the validation set, synthesis errors concentrate in three regions: 1. Tumor cores: As discussed above, enhancement unpredictability leads to under-detailed cores. 2. Skull boundaries: The sharp intensity transition at the skull-brain interface occasionally produces ringing artifacts in synthesized T1/T1c. 3. CSF spaces: FLAIR synthesis shows the highest error in CSF-containing regions (ventricles, sulci), where the CSF suppression effect is difficult to predict precisely from T1/T2 contrast.

Flow Matching vs. Alternative Formulations Several recent methods exploit informative priors for efficient generation. I^2SB (Liu et al., 2023) formulates image-to-image translation as a Schrödinger bridge between source and target distributions. BBDM (Li et al., 2023) uses Brownian bridge processes for the same purpose. LBM (Chadebec et al., 2025) operates in latent space with bridge matching. How does WFM relate to these approaches?

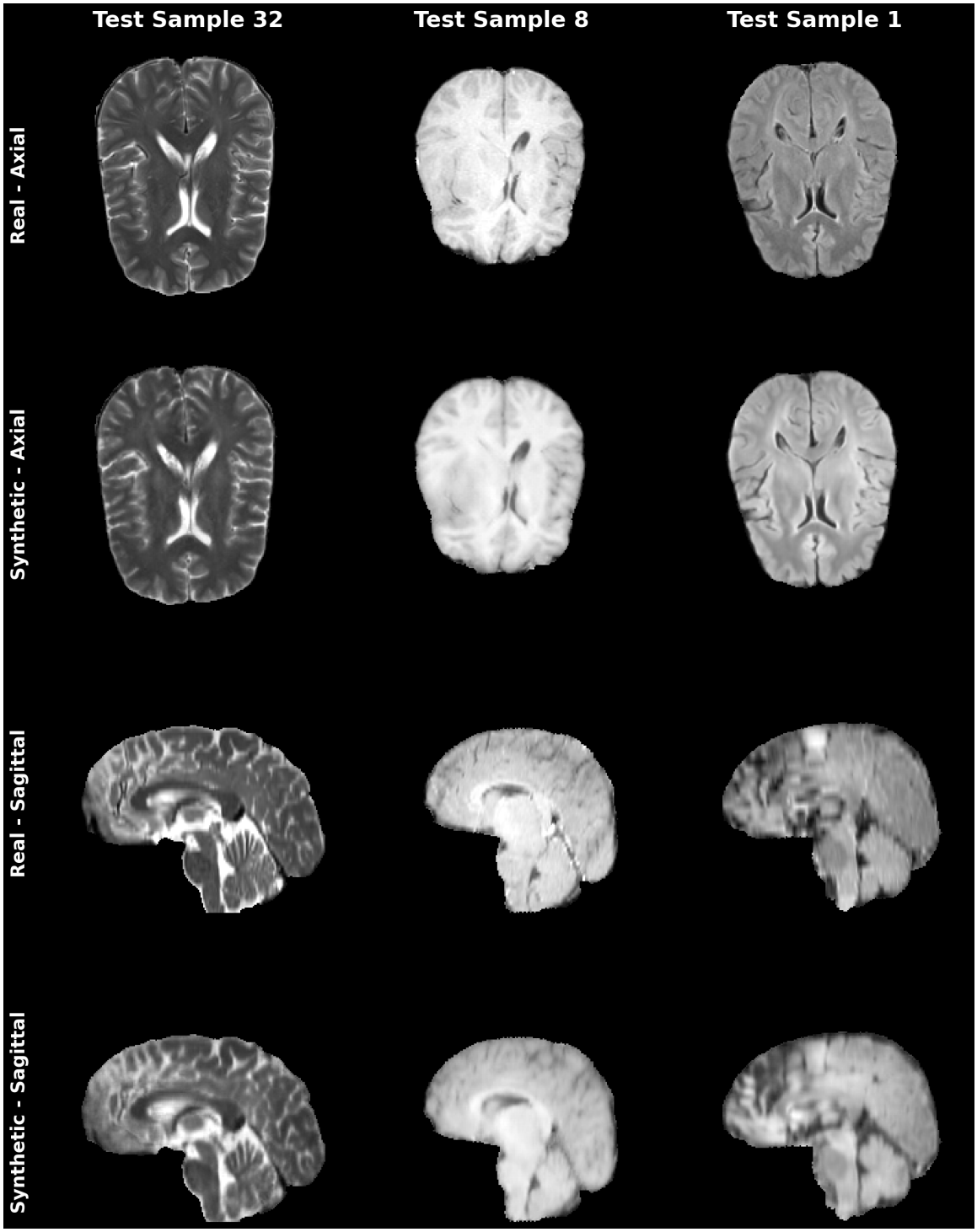

Figure 3: Additional qualitative results (samples 32, 8, 1). The method generalizes across the validation set while preserving structural details.

The key distinction is simplicity. Flow matching with a linear interpolation path yields a constant velocity field, enabling single-step inference without complex ODE solvers or bridge-specific training objectives. While I^2SB and BBDM achieve strong results, they typically require 10-50 steps for optimal quality. WFM trades theoretical generality for practical efficiency: by assuming that source and target are related by a simple contrast transformation, we obtain a formulation where one-step integration is sufficient.

This assumption is well-matched to multi-modal MRI synthesis but may not hold for more distant translation tasks (e.g., CT-to-MRI, where anatomical correspondence is weaker). For such settings, bridge-based methods with their greater flexibility may be preferable.

Wavelet vs. Latent Space Processing WFM operates in wavelet space rather than the latent space used by methods like Adaptive Latent Diffusion (Kim and Park, 2024). This choice has specific trade-offs.

Wavelet decomposition is invertible and deterministic: the 3D Haar transform compresses spatial dimensions by 2× in each axis while preserving all information across eight subbands. There is no learned encoder, no reconstruction loss, and no risk of information bottleneck. The model operates directly on image content, not on a learned abstraction.

Latent diffusion, by contrast, uses a pretrained autoencoder to compress images into a lower-dimensional space. This can yield greater compression (8× or more) but introduces reconstruction error and requires the autoencoder to generalize across the target domain. For medical imaging, where subtle lesion details matter, wavelet space provides a more conservative choice that guarantees lossless reconstruction.

The 4x parameter reduction in WFM (82M vs. 326M for four cWDM models) comes not from aggressive compression but from weight sharing across modalities—a benefit of the unified architecture rather than the wavelet representation itself.

The Quality-Speed Trade-off WFM achieves 26.8 dB average PSNR compared to cWDM's 28.4 dB—a gap of approximately 1.6 dB. Is this trade-off acceptable?

The answer depends on the application. For real-time clinical review, where a radiologist needs to visualize a missing modality during a reading session, sub-second synthesis with 0.94 SSIM may be preferable to waiting 160 seconds for marginally higher fidelity. For research applications requiring maximum accuracy—such as training segmentation models on synthetic data—the quality gap may matter more.

Notably, the gap is not uniform across modalities. T1 synthesis (27.6 dB) approaches cWDM quality (29.7 dB), while FLAIR shows a larger gap (26.1 vs. 27.8 dB). This modality dependence likely reflects the varying difficulty of contrast prediction: FLAIR's CSF suppression creates intensity patterns that are less predictable from T1/T2/T1c combinations.

Additional limitations include:

- **Single dataset evaluation.** All experiments use BraTS 2024, which consists of glioma patients. Generalization to other pathologies (metastases, stroke, multiple sclerosis) and healthy anatomy remains unvalidated.

- **No downstream task evaluation.** We report PSNR and SSIM but do not evaluate whether synthetic modalities improve tumor segmentation accuracy. This downstream validation is essential for clinical utility claims. In the future work, we will conduct segmentation as a downstream task.

- **Fixed conditioning configuration.** WFM assumes exactly three conditioning modalities are available. Handling variable numbers of inputs (one, two, or three available sequences) would require architectural modifications.

- **Comparison limited to diffusion.** We benchmark against cWDM but not against GAN-based methods (Pix2Pix, CycleGAN, ResViT) or other flow-based approaches. A broader comparison would strengthen the efficiency claims.

Clinical Deployment Considerations The 250-1000x speedup has practical implications beyond raw throughput. At 0.16-0.64 seconds per volume, WFM can run on-demand during clinical sessions rather than requiring batch processing. This enables interactive workflows: a radiologist reviewing an incomplete study could request synthesis of missing modalities in real time.

However, clinical deployment raises additional considerations not addressed in this work. Regulatory approval requires extensive validation across diverse patient populations. Uncertainty quantification—knowing when a synthesis is unreliable—is essential for safe clinical use. And integration with clinical PACS systems demands engineering effort beyond the algorithmic contribution.

We view WFM as a step toward clinically viable synthesis rather than a complete solution. The speed barrier has been addressed; the validation and integration challenges remain.