SPATIAL–FREQUENCY-GUIDED PROGRESSIVE FLOW MATCHING FOR ACCURATE 3T-TO-7T MRI SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthesizing 7T-like images from 3T MRI enhances anatomical clarity and expands the utility of existing 3T datasets. However, since fine tissue details and high-frequency noise occupy similar spectral bands, unconstrained high-frequency enhancement tends to amplify noise and introduce artifacts, whereas excessive smoothing obscures delicate boundaries. Consequently, leveraging source image structures to constrain local detail enhancement remains a major challenge in cross-field synthesis. To address this, we propose a spatio-frequency-guided progressive conditional flow matching framework. By integrating spatial and frequency-domain representations, our approach employs a multi-band generation strategy to schedule updates across distinct frequency bands. This conditions local detail refinement directly on the source structure, enabling adaptive, sequential enhancement. Experiments on paired datasets demonstrate that our method achieves superior synthesize performance in just 5 inference steps, significantly improving pixel-level reconstruction quality and reducing high-frequency error relative to the standard flow-matching baseline.

1 INTRODUCTION

Ultra-high-field 7-Tesla (7T) magnetic resonance imaging (MRI) provides superior tissue contrast and structural resolution (Wang et al., 2021; Okromelidze et al., 2023; Welton et al., 2023; Perera Molligoda Arachchige & Garner, 2023; Hütter et al., 2020). However, prohibitive equipment costs and limited accessibility severely restrict its widespread clinical adoption. To expand the utility of routine 3T scans, researchers have increasingly explored cross-field image synthesis. Synthesizing 7T-equivalent images directly from 3T data offers a practical pathway for fine-grained brain structural analysis using existing clinical datasets (Bahrami et al., 2016; Qu et al., 2020).

Researchers have explored various methods for cross-field image synthesis, including conditional generative adversarial networks (Isola et al., 2017; Dalmaz et al., 2022), hybrid architectures, and conditional diffusion models (Özbey et al., 2023; Arslan et al., 2025). Unlike open-domain generation, which prioritizes sample diversity and perceptual realism, paired medical image translation strictly requires preserving anatomical correspondence with the source subject. Specifically, 3T-to-7T synthesis must enhance target-domain appearance while maintaining exact tissue localization and structural integrity. Flow matching addresses this by directly connecting paired source and target images through continuous, velocity-driven trajectories. This incorporates paired relationships directly into path construction, enabling explicit control over intermediate states and update schedules (Lipman et al., 2023; Tong et al., 2024). Initialized from observed source images, this progressive refinement paradigm aligns naturally with the demands of medical image translation for structural fidelity and process control. Consequently, flow matching has gained increasing adoption in this domain (Imre et al., 2026; Lin et al., 2026).

Although current flow-matching methods for medical image translation (Imre et al., 2026; Lin et al., 2026) successfully capture target 7T appearance characteristics, accurately recovering fine anatomical boundaries and local structures remains plagued by the dual challenges of oversmoothing and distortion. This issue arises because high-frequency spectral responses, which generated by both genuine tissue boundaries and high-frequency noise, do not directly equate to meaningful structural

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

content. Consequently, response magnitude alone cannot determine whether high-frequency details conform to the true anatomical morphology of the source scan. As a result, conventional detail smoothing inadvertently attenuates valuable fine structures, whereas unconstrained high-frequency enhancement introduces textural artifacts, ultimately impeding the synthesis of details faithful to individual subject anatomy.

To address these challenges, we intorduce a new 3T-to-7T MRI synthesis framework grounded in spatio-frequency synergy and progressive flow matching, which integrates multi-scale detail updates with explicit structural guidance from the source image. First, we propose a multi-band progressive trajectory that schedules the evolution of distinct frequency bands, gradually increasing the contribution of mid-to-high-frequency generation across subsequent inference steps. To ensure generated details strictly align with the underlying anatomy, we jointly leverage spatial and multi-scale frequency representations from the source image. The former captures anatomical morphology and boundary locations, whereas the later characterizes local scale-dependent detail variations. Our model fuses these source representations with currently generated features to estimate local structural support, dynamically modulating the amplitude of mid-to-high-frequency refinement via a soft gating mechanism. Finally, we apply an evidence-aware band-wise supervision strategy to the inference-consistent, unrolled few-step outputs, simultaneously enforcing structural correspondence and suppressing spurious high-frequency noise.The main contributions of this paper are summarized as follows:

- We construct a multi-band flow-matching architecture featuring a band-wise residual refinement branch, utilizing differentiated evolution schedules to explicitly organize multi-scale feature dynamics during few-step synthesis.
- We propose a spatio-frequency guidance algorithm that fuses spatial and multi-scale frequency information from the source scan to quantify local structural support, dynamically modulating mid-to-high-frequency refinement through soft gating.
- We design an evidence-aware multi-band endpoint supervision mechanism that constrains inference-consistent unrolled trajectories by uniting regional and band-wise target supervision with local velocity learning.

Extensive experiments demonstrate that our 5-step inference framework achieves the highest PSNR among competing methods and reduces HFEN by 4.4% over RFlow, effectively balancing high-frequency detail enhancement with textural artifact suppression.

2 BACKGROUND

This section positions our approach relative to existing methods and introduces the notation used throughout the paper.

2.1 RELATED METHODS

Flow-based MRI synthesis can incorporate structural constraints, such as tumor-segmentation supervision (Chang et al., 2025). K-Flow and DFM organize generation across scales (Du et al., 2025; Haji-Ali et al., 2025). Frequency-based losses emphasize difficult-to-reconstruct components (Jiang et al., 2021) or regions with high local wavelet energy (Sigillo et al., 2026). However, reconstruction difficulty and frequency energy do not directly establish anatomical correspondence with the source image. Our method combines multiband scheduling with local source-guided refinement and supervises the complete few-step rollout. A detailed discussion of related work is provided in Appendix A.

2.2 PAIRED CONDITIONAL FLOW MATCHING

Conditional flow matching achieves a continuous transformation from the source distribution to the target distribution by learning the velocity field along a predefined trajectory. In paired cross-field synthesis, let x denote the 3T source image and y denote the corresponding 7T reference image. Initialized from the source image, the generation process is defined by an ordinary differential equation

(ODE):

$$\frac{\mathrm{d}\phi_t}{\mathrm{d}t} = v_\theta(\phi_t, t; x), \quad \phi_0 = x, \tag{1}$$

where ϕ_t denotes the intermediate image during generation, and v_θ represents the velocity field predicted by the neural network.

To construct training supervision using paired data, we employ noise-free linear interpolation to define the reference trajectory $z_t = (1-t)x + ty$, with its target velocity given by $u_t = y - x$. The fundamental training objective for the model to predict this velocity is

$$L_{\mathrm{vel}} = \mathbb{E}\big[\big\|v_\theta(z_t, t; x) - u_t\big\|_2^2\big]. \tag{2}$$

2.3 Multi-scale Image Representation

Multi-scale representation separates coarse-scale information from fine-scale variations in an image through smoothing operations of varying degrees. Let G_s and G_l denote linear smoothing operators with small and large window sizes, respectively. An image I can be decomposed into low-frequency (B_L), mid-frequency (B_M), and high-frequency (B_H) components:

$$B_L(I) = G_l(I), \quad B_M(I) = G_s(I) - G_l(I), \quad B_H(I) = I - G_s(I). \tag{3}$$

These three components sum to the original image, individually characterizing local variations across different frequency bands. In cross-field synthesis, since both tissue boundaries and noise can produce strong high-frequency responses, identical high-frequency magnitudes do not guarantee consistency in local anatomical morphology. Consequently, these frequency components must be processed in conjunction with spatial structural correspondence.

2.4 Finite-Step Generation and Truncation Error

During inference, the continuous generation process is solved through numerical integration. Employing an N-step uniform Euler update, let the step size be $h = 1/N$ and $t_k = kh$. The discrete update rule is given by

$$z_{k+1} = z_k + h\, v_\theta(z_k, t_k; x), \tag{4}$$

where $z_0 = x$, and the final output is $\hat{y} = z_N$. Since each step takes the previously generated image as input, the truncation error of discrete integration and the network prediction error accumulate over successive steps, degrading the final image quality. Targeting this few-step inference process, subsequent sections directly apply supervision to the unrolled endpoint.

3 Methodology

We propose a conditional flow matching framework initialized from the 3T source image to generate 7T-like images within finite steps (Figure 1). First, we design a multi-band progressive trajectory to schedule the evolution across different scales (Section 3.1). Subsequently, fusing the spatial and frequency information of the source image, we modulate the amplitude of additional detail refinement at each step (Section 3.2). During training, we apply evidence-aware band-wise supervision to the complete unrolled results, simultaneously constraining detail correspondence and excessive high-frequency responses (Section 3.3).

3.1 Multi-band Progressive Flow Trajectory

The standard linear trajectory employs a uniform interpolation schedule to advance all frequency bands synchronously. Considering that the 3T source image already provides the primary anatomical structure, we adopt a coarse-to-fine progressive schedule. This allows detail refinement to focus more on intermediate states that have completed partial coarse-scale translation. Specifically, low-frequency variations span the entire process, while mid-to-high-frequency variations gradually concentrate in subsequent steps. Simultaneously, by modulating additional mid-to-high-frequency refinement with time gates consistent with the trajectory schedule, the network updates align with the reference trajectory. Compared to merely increasing the loss weight for mid-to-high frequencies,

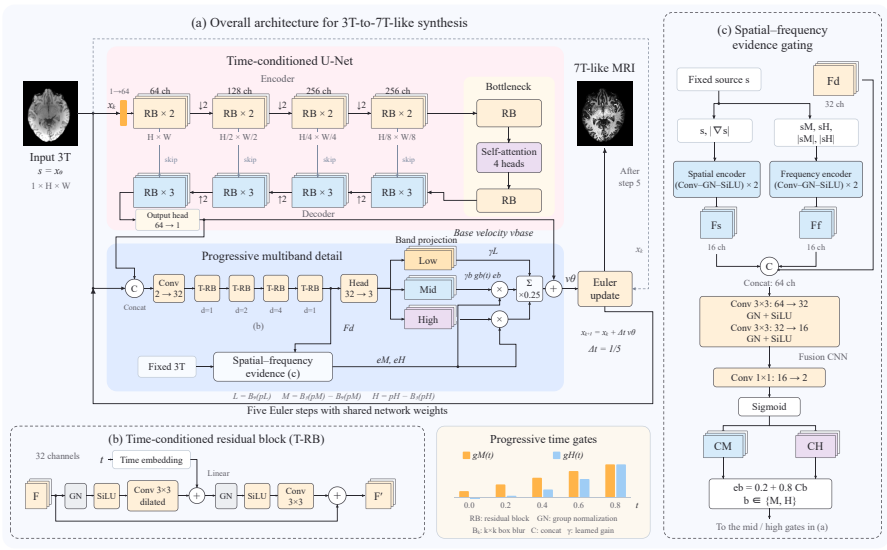


Figure 1: Overview of the proposed framework.

this design directly adjusts the update timing across different scales, explicitly incorporating the detail refinement requirements at each stage into trajectory supervision and network computation.

Following the band-wise decomposition in Section 3.2, let B_b denote the frequency band operator, where $b \in \{L, M, H\}$. We assign an independent cumulative schedule $a_b(t)$ for each frequency band to construct the training reference trajectory:

$$z_t = \sum_{b\in\{L,M,H\}} \big[B_b(x) + a_b(t)\big(B_b(y) - B_b(x)\big)\big]. \tag{5}$$

Here, x and y represent the paired 3T source image and 7T reference image, respectively. Each schedule function satisfies $a_b(0) = 0$ and $a_b(1) = 1$, ensuring that the trajectory still starts at x and ends at y. This design maintains the total variation of each frequency band unchanged, only adjusting the temporal allocation of these variations during the generation process.

For the five-step generation adopted in this paper, the variation proportions for low, mid, and high frequencies within each interval are set as follows:

$$q_L = \frac{(1,1,1,1,1)}{5}, \quad q_M = \frac{(1,2,3,4,5)}{15}, \quad q_H = \frac{(0,1,4,9,16)}{30}. \tag{6}$$

The sum of proportions in each group is 1. The cumulative schedule $a_b(t)$ utilizes linear interpolation within each interval. Letting $N = 5$, the derivative in the k-th interval is $a'_b(t) = N q_{b,k}$, and the corresponding target velocity is

$$u_t = \sum_{b\in\{L,M,H\}} a'_b(t)\big[B_b(y) - B_b(x)\big]. \tag{7}$$

Consequently, the low-frequency component maintains a uniform variation across steps, while the variation proportions for mid-to-high frequencies gradually increase. By altering the reference states and velocity targets, this arrangement explicitly incorporates the coarse-to-fine update principle into flow matching training.

To coordinate the additional refinement within the network with the reference trajectory, we establish low, mid, and high-frequency residual branches alongside the base velocity field. The base velocity field continues to predict the overall image variation, while each branch provides additional refinement at its corresponding scale. The mid-to-high-frequency branches employ time gates consistent with the trajectory change rates:

$$g_b(t) = \frac{q_{b,k}}{\max_j q_{b,j}}, \quad b \in \{M, H\}. \tag{8}$$

The time gates enable the additional mid-to-high-frequency refinement to increasingly contribute to subsequent updates. While they control the temporal schedule of the refinement, the refinement amplitude at specific locations still depends on the structural evidence provided by the source image. Therefore, in the next section, we further integrate the spatial and frequency information of the source image to locally modulate the additional mid-to-high-frequency refinement.

3.2 Spatio-Frequency-Guided Learning for Detail Support

The progressive trajectory schedules the update timing across different scales, but the amplitude of local refinement still depends on the structural evidence provided by the source image. Since both tissue boundaries and noise can produce strong high-frequency responses, relying solely on response magnitudes makes it difficult to determine the appropriate amount of detail refinement to retain. Therefore, we integrate the spatial structure and frequency band information of the source image to learn local detail support, utilizing it to modulate the amplitude of additional mid-to-high-frequency refinement.

The spatial branch takes the source image and its gradient magnitude $[x, |\nabla x|]$ as input to extract tissue appearance and boundary cues. The frequency branch takes $[B_M(x), B_H(x), |B_M(x)|, |B_H(x)|]$ as input, characterizing variation patterns and response magnitudes across different scales. These two types of features jointly provide the source image reference. Because the gating mechanism modulates the refinement proposed at the current step, we further fuse them with the current detail branch features to predict the mid-frequency and high-frequency support, denoted as C_M and C_H. Consequently, the support prediction simultaneously utilizes the structural cues from the fixed source image and the current generation state.

To ensure that the predicted support reflects the local correspondence between the source image and target details, we construct supervision targets based on the gradient and frequency consistency of the paired 3T and 7T images, training the support prediction branch accordingly. Within a local neighborhood, by integrating the consistency of variation patterns and response magnitudes, we calculate the gradient consistency A_S, mid-frequency consistency A_M, and high-frequency consistency A_H, respectively (Appendix B.1). We define the evidence as

$$E_M = 0.6A_M + 0.4A_S, \quad E_H = 0.7A_H + 0.3A_S. \tag{9}$$

We employ mean squared error to supervise C_M and C_H to approximate E_M and E_H, respectively, ensuring that the predicted support is grounded in the local spatio-frequency consistency of the paired images. In this way, the learning of the gating mechanism possesses a clear structural reference, rather than freely adjusting feature amplitudes during the reconstruction process.

Weak support from the source image does not imply the absence of target details. Some structures appear faint in 3T images, and completely closing the refinement pathway would restrict these regions from utilizing target-domain priors for refinement. Therefore, we employ a soft gating mechanism with a non-zero lower bound:

$$h_b = c_{\min} + (1 - c_{\min})C_b, \quad b \in \{M, H\}, \tag{10}$$

where $c_{\min} = 0.2$. A higher support retains a larger amplitude of additional refinement, whereas a lower support reduces the refinement amplitude while preserving the refinement pathway.

Incorporating the residual amplitude coefficients into the band-wise refinement r_b, the complete velocity field can be abbreviated as

$$v_\theta = v_{\text{base}} + r_L + g_M(t)\big(h_M \odot r_M\big) + g_H(t)\big(h_H \odot r_H\big), \tag{11}$$

where g_M and g_H are the time gates from Section 3.1, and $\odot$ denotes element-wise multiplication. The time gates schedule the stage-wise contribution of additional refinement, while the support gates modulate its amplitude based on local source image cues. The paired 7T images are utilized strictly for training supervision; during inference, the support is predicted solely from the source 3T image and the current generation features.

3.3 Multi-band Endpoint Constraints

Local fitting cannot directly guarantee the output quality after sequential unrolling. Therefore, initialized from the 3T source image, we perform a complete five-step unrolling consistent with the inference process to obtain the output $\hat{y}$, and apply multi-band endpoint supervision to it.

The supervision weights employ the fixed evidence E_M and E_H constructed in Section 3.2, preventing the predicted support from circumventing the penalty by directly altering the loss weights. The low-frequency loss L_L utilizes the Charbonnier penalty (Charbonnier et al., 1997; Lai et al., 2017), a smooth approximation of the absolute error, to constrain overall intensity and large-scale variations. The mid-frequency loss L_M weights the pixel-wise reconstruction error with E_M and the Sobel gradient difference with $1 - E_M$, shifting the supervision focus from numerical alignment to boundary variations as local evidence decreases.

The high-frequency supervision further distinguishes between detail positional correspondence and local magnitude. We define

$$R(I) = G_7\big(B_H(I)^2\big) + \varepsilon^2, \tag{12}$$

where G_7 denotes the 7×7 local spatial average, and $R(I)$ measures the local high-frequency magnitude. When the evidence is strong, the alignment term $L_{H,\text{align}}$ constrains the pixel-wise Charbonnier error of the high-frequency component, emphasizing detail correspondence. When the evidence is moderate, the magnitude matching term $L_{H,\text{rms}}$ compares $R(\hat{y})$ with $R(y)$, constraining the detail magnitude within the neighborhood without demanding strict pixel-wise consistency. When the evidence is weak, the excess suppression term $L_{H,\text{excess}}$ only penalizes the predicted magnitude that exceeds $(1 + \delta)R(y) + \varepsilon$, preventing the network from forcefully fabricating target high frequencies in regions lacking sufficient source image evidence.

These three terms are continuously weighted by E_H^2, $2E_H(1 - E_H)$, and $(1 - E_H)^2$, respectively, formulating the overall high-frequency loss:

$$L_H = L_{H,\text{align}} + 0.5\, L_{H,\text{rms}} + L_{H,\text{excess}}. \tag{13}$$

The terminal loss is given by

$$L_{\text{terminal}} = L_L + 2L_M + 4L_H, \tag{14}$$

where $\delta = 0.1$. Each term employs a uniform spatially weighted average $\langle\cdot\rangle_w$, assigning higher weights to intra-brain regions while retaining smaller weights for the background (Appendix B.1).

The final training objective is

$$L_{\text{total}} = L_{\text{vel}} + L_{\text{local}} + 0.1\, L_{\text{conf}} + L_{\text{terminal}}, \tag{15}$$

where L_{local} supervises the subsequent reference state via image reconstruction, SSIM, gradient, and band-wise errors, and L_{conf} supervises the support prediction. While the local loss learns the updates at each step, the terminal loss directly constrains the result obtained by executing these updates sequentially.

4 Experiments

4.1 Datasets and Experimental Setup

Datasets and preprocessing. We use paired 3T–7T T1-weighted MRI from 80 subjects: 10 healthy adults from the UNC dataset (Chen et al., 2023), 20 from the multimodal 3T–7T dataset (Li et al., 2024), and 25 from each of two de-identified internal clinical cohorts. The training, validation, and test sets contain 56, 8, and 16 subjects (4,800, 675, and 1,200 slices), respectively, with no subject overlap. Further dataset details are provided in Appendix B.3.

All raw data are 3D volumes in NIfTI format and undergo a unified preprocessing pipeline. First, brain tissue masks are generated using FreeSurfer (Fischl, 2012; Ségonne et al., 2004) to remove extra-cranial regions. Subsequently, designating the 7T images as fixed references, the corresponding 3T images are registered and resampled to the 7T space via FSL FLIRT (Jenkinson et al., 2002; Jenkinson & Smith, 2001). This registration process employs a 12-degree-of-freedom affine transformation with normalized mutual information as the cost function, applying trilinear interpolation. After registration, the volumetric data are converted into 2D slices. For each slice, the intensities are clipped to the 0.3rd and 99.7th percentiles and subsequently saved as 8-bit grayscale PNG images.

Implementation details. All experiments use 2D single-channel inputs of 320×256. We initialize the backbone from the best base flow-matching checkpoint and jointly train the model with AdamW

Table 1: Quantitative comparison on the test set. All methods are evaluated under the same protocol, identical test samples, and identical intensity ranges. Best value per metric is shown in bold; ↑ (↓) indicates that higher (lower) is better.

Method	PSNR↑	SSIM↑	HFEN↓	LPIPS↓	MAE↓
pix2pix	22.059	0.816	0.756	0.097	0.0919
ResViT	22.974	0.840	0.675	**0.085**	0.0824
WATNet	21.410	0.792	0.835	0.184	0.0983
VNet	23.449	0.8605	0.691	0.171	0.0811
GDiff	23.362	**0.8606**	**0.636**	0.132	0.0899
RFlow	23.178	0.846	0.697	0.096	0.0788
SynDiff	23.497	0.831	0.741	0.158	0.0857
SelfRDB	23.592	0.845	0.662	0.195	**0.0763**
Ours	**23.970**	0.847	0.666	0.118	0.0788

on two NVIDIA RTX 3090 GPUs. Training uses a base learning rate of 1×10^{-4} and an effective batch size of eight slices; checkpoints are selected by validation total loss. Both training and inference use five Euler steps. Detailed network configurations and training settings are provided in Appendices B.1 and B.4.

Evaluation metrics. We employ PSNR, SSIM (Wang et al., 2004), HFEN, LPIPS (Zhang et al., 2018), and the brain tissue foreground MAE to evaluate the generated results in terms of pixel-level reconstruction, structural consistency, high-frequency errors, and perceptual differences. All methods are compared under identical test samples, image intensity ranges, and spatial processing protocols to ensure a consistent evaluation standard.

4.2 Main Results

We compare the proposed method with several representative 3T-to-7T synthesis and medical image translation approaches, including pix2pix (Isola et al., 2017), GDiff (Dhariwal & Nichol, 2021), ResViT (Dalmaz et al., 2022), VNet (Milletari et al., 2016), WATNet (Qu et al., 2020), SynDiff (Özbey et al., 2023), SelfRDB (Arslan et al., 2025), and Rectified Flow (RFlow) (Liu et al., 2023). All methods are evaluated on the same testing set under a unified evaluation protocol, with quantitative results presented in Table 1.

In terms of overall performance, our method achieves the highest PSNR and yields further improvements over the base flow matching model. This demonstrates that the proposed multi-band progressive restoration and source image evidence constraints effectively enhance the overall pixel-level reconstruction quality. Our method obtains these results using only five velocity field evaluations, indicating that the proposed design is well-suited for finite-step generation settings.

Figure 2 further provides a visual comparison across different methods. We select brain tissue boundaries, fine anatomical structures, and local texture regions for magnified visualization. Identical intensity ranges and crop coordinates are applied to all methods, enabling a comparative analysis of how different models restore local details beyond overall reconstruction.

4.3 Ablation Study

To analyze the contribution of each component, we apply two inference-stage interventions on the same checkpoint. The `no_detail` variant deactivates the additional detail pathway, whereas the `no_evidence` variant sets the mid-to-high-frequency effective gates to 1. All variants maintain identical five-step inference and evaluation settings. The results are presented in Figure 3 and Table 2.

Upon deactivating the detail pathway, LPIPS increases by approximately 25.4%, yielding smoother local structures in the generated images. Meanwhile, SSIM and HFEN show slight improvements, with MAE remaining largely unchanged. This observation indicates that the additional detail refine-

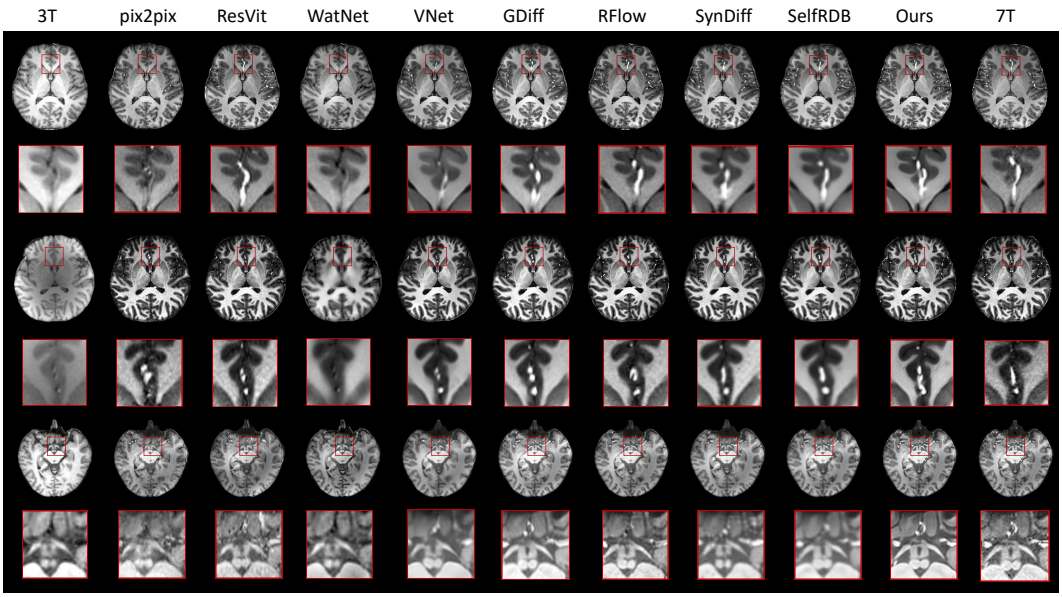


Figure 2: Visual comparison of 3T-to-7T synthesis results. Magnified regions cover tissue boundaries, fine anatomical structures, and local texture. All methods are displayed with identical intensity ranges and crop coordinates.

Table 2: Inference-stage interventions applied to the same trained checkpoint. All variants use identical five-step inference and evaluation settings; none of them is retrained. Best value per metric is shown in bold.

Variant	PSNR↑	SSIM↑	HFEN↓	LPIPS↓	MAE↓
Ours	**23.970**	0.847	0.666	**0.118**	**0.0788**
no_detail	23.895	**0.853**	**0.655**	0.148	**0.0788**
no_evidence	22.891	0.824	0.766	0.127	0.0867

ment substantially influences local image representation, although its benefits do not manifest as a simultaneous improvement across all evaluation metrics.

Removing the gating attenuation results in a PSNR drop of approximately 1.08 dB, accompanied by a degradation in all other metrics. Visual comparisons reveal that this variant produces excessively strong texture responses and local regular fine lines. These results demonstrate that the additional detail refinement in the current model necessitates appropriate amplitude control.

4.4 Downstream Tasks

To evaluate whether the structural information in the generated images can benefit downstream automated analysis, we employ a fixed SynthSeg model to segment the hippocampal regions on 3T, generated 7T-like, and real 7T images from two test subjects for comparison. The experiment comprises two comparison setups. The first setup maps manual T2 labels to the T1 space and utilizes the manual reference to assist in centroid translation for alignment, thereby evaluating shape consistency after position correction. The second setup directly uses the automated segmentation on real 7T images as the reference without translation.

We employ the Dice score, recall, precision, and relative volume error to evaluate the regional overlap, coverage extent, and volume consistency of the hippocampal segmentations, with the resulting metrics presented in Table 3 and Table 4.

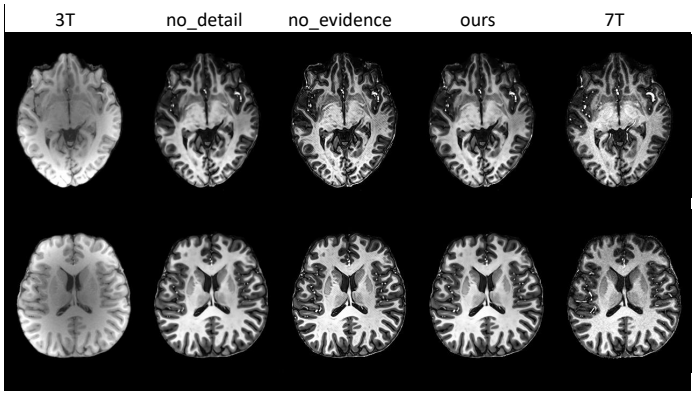


Figure 3: Visual comparison of the inference-stage interventions on the same checkpoint. All variants use identical five-step inference and evaluation settings.

Table 3: Hippocampal segmentation consistency under the manual-reference setup. Manual T2 labels are mapped to the T1 space and used to assist centroid translation before evaluation. Recall, precision, and volume error are reported in percent.

Input image	Dice↑	Recall (%)↑	Precision (%)↑	Volume error (%)↓
Original 3T	0.6503	76.76	56.45	**35.26**
Synthesized 7T-like	0.6589	79.38	56.34	40.43
Real 7T	**0.6642**	**80.09**	**56.75**	40.65

Table 4: Hippocampal segmentation consistency under the real-7T setup. The automatic SynthSeg segmentation of the real 7T images is used directly as the reference, without centroid translation. Recall, precision, and volume error are reported in percent.

Input image	Dice↑	Recall (%)↑	Precision (%)↑	Volume error (%)↓
Original 3T	0.9058	88.84	**92.40**	3.86
Synthesized 7T-like	**0.9217**	**92.09**	92.24	**0.17**

Under the manual reference setup, the generated images exhibit slight improvements in Dice and recall, but an increase in volume error. When referencing the automated segmentation of real 7T images, the Dice scores for both subjects improve, recall increases by approximately 3.26 percentage points, volume error decreases, and precision remains comparable. These results tentatively support that the generated images provide segmentable structures closer to those of 7T images. However, constrained by the limited sample size and evaluation methodology, we cannot yet definitively assert the absolute authenticity of the generated details.

5 Conclusion

We propose a spatio-frequency-guided progressive flow matching framework for 3T-to-7T MRI synthesis to address detail smoothing and textural artifacts. It explicitly organizes multi-scale feature evolution via a multi-band progressive trajectory and adaptively modulates local detail refinement using source spatial and frequency information. By incorporating evidence-aware multi-band endpoint constraints, this framework synergistically optimizes overall appearance and local details within finite steps. Experiments demonstrate our five-step inference method significantly outperforms representative approaches in pixel reconstruction and high-frequency error control. Furthermore, downstream hippocampus segmentation confirms the generated images provide segmentable structures akin to real 7T data, expanding the utility of routine 3T scans. Future work will explore reliably utilizing target-domain priors to ensure anatomical credibility in regions lacking sufficient source structural cues.

AI USE STATEMENT

We used generative AI tools to assist with developing the conceptual framework, writing proof-related material, interpreting experimental results, and translating parts of the manuscript. We also used generative AI tools to draft and edit portions of the paper for readability, identify relevant literature, and format references.

We did not use generative AI tools to generate synthetic datasets, formulate mathematical claims or hypotheses,provide feedback on the research methodology and experimental design, provide critical ingredients for proofs, implement the proposed method, clean or reformat datasets, support qualitative or thematic analysis, design surveys or interviews, or create or modify scientific figures or images.

All AI-assisted material was reviewed and verified by the authors. The authors checked the mathematical expressions, citations, methodological descriptions, experimental interpretations, and final wording against the research records and source materials. The authors take full responsibility for the content of this submission, including all claims, analyses, and artifacts produced with the assistance of generative AI tools.

REFERENCES

Fuat Arslan, Bilal Kabas, Onat Dalmaz, Muzaffer Özbey, and Tolga Çukur. Self-consistent recursive diffusion bridge for medical image translation. *Medical Image Analysis*, 106:103747, 2025. doi: 10.1016/j.media.2025.103747.

Khosro Bahrami, Feng Shi, Xiaopeng Zong, Hae Won Shin, Hongyu An, and Dinggang Shen. Reconstruction of 7T-Like images from 3T MRI. *IEEE Transactions on Medical Imaging*, 35 (9):2085–2097, 2016. doi: 10.1109/TMI.2016.2549918.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models, 2025.

Heng Chang, Yu Shang, Haifeng Wang, Yuxia Liang, Haoyu Wang, Fan Wang, Chen Niu, and Chunfeng Lian. Controllable flow matching for 3D contrast-enhanced brain MRI synthesis from non-contrast scans. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15975 of *Lecture Notes in Computer Science*, pp. 119–128. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05325-1_12. URL https://papers.miccai.org/miccai-2025/0180-Paper1113.html.

Pierre Charbonnier, Laure Blanc-Feraud, Gilles Aubert, and Michel Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 6(2): 298–311, 1997. doi: 10.1109/83.551699.

Xiaoyang Chen, Liangqiong Qu, Yifang Xie, Sahar Ahmad, and Pew-Thian Yap. A paired dataset of T1- and T2-weighted MRI at 3 tesla and 7 tesla. Figshare, 2023. Collection.

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. doi: 10.1109/TMI.2022.3167808.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 8780–8794, 2021.

Weitao Du, Shuning Chang, Jiasheng Tang, Yu Rong, Fan Wang, and Shengchao Liu. Flow along the K-amplitude for generative modeling, 2025.

Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

Moayed Haji-Ali, Willi Menapace, Ivan Skorokhodov, Arpit Sahni, Sergey Tulyakov, Vicente Ordonez, and Aliaksandr Siarohin. Improving progressive generation with decomposable flow matching, 2025.

Bernd-Otto Hütter, Jan Altmeppen, Oliver Kraff, Stefan Maderwald, Jens M. Theysohn, Adrian Ringelstein, Karsten H. Wrede, Philipp Dammann, Harald H. Quick, Marc Schlamann, and Christoph Moenninghoff. Higher sensitivity for traumatic cerebral microbleeds at 7 T ultra-high field MRI: is it clinically significant for the acute state of the patients and later quality of life? *Therapeutic Advances in Neurological Disorders*, 13:1756286420911295, 2020. doi: 10.1177/1756286420911295.

Baris Imre, Aram Salehi, Levente Baljer, Andrew Webb, Marius Staring, and Efe Ilicak. Conditional flow matching for cross-field MRI harmonisation, 2026.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, 2017.

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001. doi: 10.1016/S1361-8415(01)00036-6.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17(2):825–841, 2002. doi: 10.1006/nimg.2002.1132.

Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13899–13909, 2021. doi: 10.1109/ICCV48922.2021.01366.

Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5835–5843, 2017. doi: 10.1109/CVPR.2017.618.

Shuyu Li, Lei Chu, Baoqiang Ma, Xiaoxi Dong, Yirong He, Debin Zeng, et al. A paired dataset of multi-modal MRI at 3 tesla and 7 tesla with manual hippocampal subfield segmentations on 7T T2-weighted images. Figshare+, 2024. Dataset.

Yiyang Lin, Chenhui Wang, Zhihao Peng, and Yixuan Yuan. UniField: A unified field-aware MRI enhancement framework, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.116.

Daniele Molino, Alessio Zoboli, Camillo Maria Caruso, Valerio Guarrasi, and Paolo Soda. Compositional cross-modality translation via whole-volume multitask latent flow matching, 2026. URL https://arxiv.org/abs/2608.08135.

Lela Okromelidze, Vishal Patel, Rahul B. Singh, Alfonso S. Lopez Chiriboga, Shengzhen Tao, Xiangzhi Zhou, Sina Straub, Erin M. Westerhold, Vivek Gupta, Amit K. Agarwal, John V. Murray, Amit Desai, S.J.S. Sandhu, I. Vanessa Marin Collazo, and Erik H. Middlebrooks. Central vein sign in multiple sclerosis: A comparison study of the diagnostic performance of 3t versus 7T MRI. *American Journal of Neuroradiology*, 45(1):76–81, 2023. doi: 10.3174/ajnr.A8083.

Muzaffer Özbey, Onat Dalmaz, Salman U. H. Dar, Hasan A. Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023. doi: 10.1109/TMI.2023.3290149.

Arosh S. Perera Molligoda Arachchige and Anton Kristoffer Garner. Seven tesla MRI in Alzheimer's disease research: State of the art and future directions: A narrative review. *AIMS Neuroscience*, 10(4):401–422, 2023. doi: 10.3934/Neuroscience.2023030.

Liangqiong Qu, Yongqin Zhang, Shuai Wang, Pew-Thian Yap, and Dinggang Shen. Synthesized 7T MRI from 3T MRI via deep learning in spatial and wavelet domains. *Medical Image Analysis*, 62:101663, 2020. doi: 10.1016/j.media.2020.101663.

F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060–1075, 2004. doi: 10.1016/j.neuroimage.2004.03.032.

Luigi Sigillo, Shengfeng He, and Danilo Comminiello. Latent wavelet diffusion for ultra-high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2026.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, volume 202, pp. 32211–32252, 2023.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research (TMLR)*, 2024.

Z. Irene Wang, Se-Hong Oh, Mark Lowe, Mykol Larvie, Paul Ruggieri, Virginia Hill, Volodymyr Statsevych, Doksu Moon, Jonathan Lee, Todd Emch, James Bena, Ingmar Blümcke, William Bingaman, Jorge A. Gonzalez-Martinez, Imad Najm, and Stephen E. Jones. Radiological and clinical value of 7T MRI for evaluating 3t-visible lesions in pharmacoresistant focal epilepsies. *Frontiers in Neurology*, 12:591586, 2021. doi: 10.3389/fneur.2021.591586.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Thomas Welton, Septian Hartono, Yao-Chia Shih, Stefan T. Schwarz, Yue Xing, Eng-King Tan, Dorothee P. Auer, Noam Harel, and Ling-Ling Chan. Ultra-high-field 7T MRI in Parkinson's disease: ready for clinical use?—a narrative review. *Quantitative Imaging in Medicine and Surgery*, 13(11):7607–7620, 2023. doi: 10.21037/qims-23-509.

Milad Yazdani, Yasamin Medghalchi, Pooria Ashrafian, Ilker Hacihaliloglu, and Dena Shahriari. Flow matching for medical image synthesis: Bridging the gap between speed and quality. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15975 of *Lecture Notes in Computer Science*, pp. 216–226. Springer Nature Switzerland, 2025. doi: 10.1007/978-3-032-05325-1_21. URL https://papers.miccai.org/miccai-2025/0343-Paper1056.html.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

APPENDIX OVERVIEW

Appendix A reviews related work. Appendix B provides implementation details, dataset descriptions, hyperparameters, and computational costs. Appendix C visualizes the predicted mid- and high-frequency support averaged over five inference steps.

A RELATED WORK

Flow matching has been adapted to medical image synthesis and translation. Yazdani et al. (2025) use optimal transport flow matching for medical generation with class- and mask-based conditioning. For paired MRI translation, Chang et al. (2025) combine a straight source-to-target path with tumor segmentation constraints to synthesize contrast-enhanced brain MRI in one step. Molino et al. (2026) combine a pretrained 3D variational autoencoder with multitask latent flow matching for whole-volume translation, including MRI-to-MRI synthesis. Our focus is complementary: we schedule refinement across frequency bands and regulate local detail updates using source spatial and frequency cues.

In cross-field synthesis and image generation tasks, multi-scale information is widely utilized for conditional guidance and generation trajectory design. For example, WATNet (Qu et al., 2020) utilizes the wavelet features of the source image to modulate spatial features, achieving 3T-to-7T image translation. Meanwhile, K-Flow (Du et al., 2025) and DFM (Haji-Ali et al., 2025) explicitly organize the progressive generation process by introducing scale parameters and scale-wise schedules, respectively. Furthermore, frequency information is frequently used to adjust training supervision signals. Focal Frequency Loss (Jiang et al., 2021) calculates the spectrum error between predicted and reference images, increasing the loss weight for hard-to-reconstruct frequency components. LWD (Sigillo et al., 2026) allocates stronger supervision to detail-rich regions based on local wavelet energy. However, these weighting mechanisms primarily rely on reconstruction errors and local energy. In paired medical image translation, high-frequency energy matching is not equivalent to local anatomical correspondence. Therefore, exploring how to incorporate spatial correspondence to guide detail generation remains necessary.

To reduce the inference steps in diffusion models and flow matching, researchers have proposed various methods to modify learning targets and training objectives. Consistency Models (Song et al., 2023) accelerate generation by learning to map different intermediate states along the same generation trajectory to a common data endpoint. Flow Map Matching (Boffi et al., 2025) further learns the mappings between any two timesteps on the generation trajectory. These works demonstrate that adjusting training objectives is an effective approach to achieving high-quality few-step generation. Whereas these methods learn endpoint or inter-time mappings, we retain velocity-field updates and directly supervise the output of sequential unrolling from the source image, optimizing reconstruction error under few-step inference.

B EXPERIMENTAL DETAILS

This appendix provides implementation and evaluation details omitted from the main paper for readability. Appendix B.1 describes the network configuration, multiband correction, local support supervision, and the five-step training and inference procedures. Appendix B.3 describes the data sources used in this study. Appendix B.2 reports parameter counts and inference costs, while Appendix B.4 summarizes the training settings, inference configuration, and model-selection protocol.

B.1 IMPLEMENTATION DETAILS

B.1.1 NETWORK CONFIGURATION AND MULTIBAND CORRECTION

We use a conditional flow-matching network as the base velocity field. The backbone is a two-dimensional, single-channel U-Net with 64 base channels and channel multipliers $(1, 2, 4, 4)$. Each resolution level contains two residual blocks, and an attention module is used at the 16×16 feature resolution. The backbone predicts the velocity of the full image.

The detail branch takes the current state and the base velocity as input. It uses 32 hidden channels and four time-conditioned residual blocks with dilation rates $(1, 2, 4, 1)$, while preserving the spatial resolution. The branch outputs three proposal maps P_L, P_M, and P_H, whose low-, mid-, and high-frequency components are used as additive corrections. The spatial-evidence and frequency-evidence encoders both use 16 hidden channels. Each encoder contains two 3×3 convolutional layers followed by group normalization and SiLU activation. The two evidence streams are fused with the current detail features, and a 1×1 convolution predicts the mid- and high-frequency support maps.

We construct the multiscale representation with stride-one mean filters. The fine and coarse windows are 3×3 and 9×9, respectively, and reflection padding preserves the spatial dimensions of all bands. For a proposal map P_b, we first extract its corresponding band and then apply a learnable amplitude factor. The additional correction is

$$r_b = 0.25\, \gamma_b B_b(P_b), \qquad \gamma_b = 2\ \mathrm{sigmoid}(\eta_b), \qquad b \in \{L, M, H\}, \tag{16}$$

where η_b is a learnable scalar initialized at zero. The mid- and high-frequency corrections are subsequently modulated by the temporal gates and support gates defined in the main paper.

B.1.2 Spatial weighting

The training images are normalized to $[-1, 1]$. We construct a foreground mask m from the union of the non-black regions in the paired images and define the spatial weight

$$w(p) = m(p) + 0.05\,[1 - m(p)]. \tag{17}$$

Foreground pixels therefore receive weight 1, while background pixels receive weight 0.05. For a pixelwise error map f, the weighted mean is

$$\langle f \rangle_w = \frac{\sum_p w(p) f(p)}{\sum_p w(p)}, \tag{18}$$

where p indexes the pixels in the current batch. This weighting is used to aggregate the training losses spatially.

B.1.3 Local agreement and support supervision

The agreement measure in Section 3.2 accounts for both local response patterns and response strength. Let s_j and r_j denote corresponding source and reference components, and let G_7 denote a 7×7 local mean filter. We define

$$\begin{aligned} c &= \sum_j G_7(s_j r_j), \\ e_s &= \sum_j G_7(s_j^2), \qquad e_r = \sum_j G_7(r_j^2), \\ q &= \sqrt{e_s e_r + \epsilon_a}. \end{aligned} \tag{19}$$

The local agreement is

$$A(s, r) = \mathrm{clip}\left(\frac{c}{q}, 0, 1\right) \mathrm{clip}\left(\frac{2q}{e_s + e_r + \epsilon_a}, 0, 1\right), \tag{20}$$

where $\epsilon_a = 10^{-8}$ and clip truncates a value to the specified interval. The first factor measures positive local response agreement, whereas the second measures agreement in response strength. Spatial agreement A_S uses horizontal and vertical Sobel gradients as the component set; mid- and high-frequency agreement, A_M and A_H, use the corresponding bands. After constructing E_M and E_H with the weights defined in the main paper, the support loss is

$$L_{\mathrm{conf}} = \frac{1}{2}\left[\left\langle (C_M - E_M)^2 \right\rangle_w + \left\langle (C_H - E_H)^2 \right\rangle_w\right]. \tag{21}$$

The targets are constructed from paired images and are treated as fixed during optimization. The 7T reference image is not required at inference time.

B.1.4 LOCAL RECONSTRUCTION SUPERVISION

During training, we sample a time point $t_k = k/5$ from the five-step grid, where $k \in \{0, 1, 2, 3, 4\}$. Given the reference state z_{t_k}, one Euler update produces

$$\widetilde{z}_{k+1} = z_{t_k} + \frac{1}{5} v_\theta(z_{t_k}, t_k; x). \tag{22}$$

We compare $\widetilde{z}_{k+1}$ with the next reference state $z_{t_{k+1}}$. The local loss is

$$L_{\text{local}} = L_{\text{rec}} + 0.2L_{\text{ssim}} + 0.1L_{\text{grad}} + 0.2L_{\text{freq}}. \tag{23}$$

Here, L_{rec} is the weighted Charbonnier loss with

$$\rho(e) = \sqrt{e^2 + \epsilon_c^2}, \qquad \epsilon_c = 10^{-3}. \tag{24}$$

L_{ssim} is the weighted $1 - \text{SSIM}$ loss, and L_{grad} is the weighted absolute difference between the horizontal and vertical Sobel gradients. The multiband reconstruction term is

$$L_{\text{freq}} = \frac{0.25\ell_L + 0.5\ell_M + \ell_H}{1.75}, \tag{25}$$

with

$$\ell_b = \left\langle \rho\big[B_b(\widetilde{z}_{k+1}) - B_b(z_{t_{k+1}})\big] \right\rangle_w. \tag{26}$$

The multiband term is normalized by the sum of its band weights. In contrast, the low-, mid-, and high-frequency terms in the terminal loss use weights 1, 2, and 4 without normalization. Multiband reconstruction and terminal high-frequency energy calculations are performed in FP32 to improve the numerical stability of small residuals.

B.1.5 TRAINING AND INFERENCE PROCEDURE

Each training iteration includes both local supervision along the reference path and supervision from a complete rollout starting at the source image. The local branch samples one reference state and learns the velocity and the next reference state. The rollout branch performs five Euler updates from the 3T source image, computes the terminal loss on the final output, and backpropagates through the complete rollout. Both branches optimize the same model.

At inference time, the state is initialized with the 3T source image. At every step, the fixed source image and the current state are used to predict the support maps, and the same network parameters are reused across all updates. The paired 7T image is used only for training supervision and is not provided to the inference procedure.

B.2 MODEL SIZE AND INFERENCE COST

Table 5 compares the parameter counts and average inference times of the evaluated methods. All methods were measured on a single NVIDIA RTX 3090 with the same input size and batch size. Parameter counts are reported in millions of parameters (M). For iterative generators, we also report the number of update steps and ensemble evaluations to make the actual generation cost explicit.

Our method reuses the same velocity-field parameters across its five updates. The paired support-target construction and terminal loss are used only during training and do not add network evaluations at inference time. Under the configurations in Table 5, our method reduces the average inference time by approximately 42.5% and 59.8% relative to SynDiff and SelfRDB, respectively. Runtime depends jointly on the network architecture, number of updates, and number of ensemble evaluations; the table therefore reports the total cost under each method's evaluation configuration rather than an architecture-only ranking.

B.3 DATASETS

We use two public datasets and two internal clinical cohorts, comprising 80 subjects with paired 3T–7T T1-weighted (T1w) MRI. The public resources contain additional modalities, but the synthesis model uses only paired T1w images. Manual T2w annotations are used for the downstream evaluation in Section 4.4 and are not used for model training.

Table 5: Parameter counts and average inference costs. Runtime was measured on one NVIDIA RTX 3090 with a common input size and batch size.

Method	Params. (M)	Inference configuration	Time (ms)
pix2pix	54.41	single forward pass	1.90
GDiff	138.28	25 steps, 5 ensembles	7525.00
ResViT	123.40	single forward pass	31.30
VNet	5.31	single forward pass	4.80
WATNet	0.63	single forward pass	22.73
Base flow matching	35.70	5 steps, 5 ensembles	1177.00
SynDiff	39.73	5 steps	208.09
SelfRDB	46.94	5 steps	297.22
Ours	24.13	5 steps	119.62

B.3.1 UNC 3T–7T PAIRED MRI DATASET

The UNC dataset contains paired T1w and T2w brain MRI from 10 healthy adults released by the University of North Carolina at Chapel Hill. The cohort includes three women and seven men aged 25–41 years. The 3T scans were acquired on a Siemens MAGNETOM Prisma system and the 7T scans on a Siemens MAGNETOM Terra system, both using 32-channel head coils. The 3T T1w and T2w scans used MPRAGE and SPACE sequences, respectively, with 0.8-mm isotropic resolution. The 7T T1w and T2w scans used MP2RAGE and TSE sequences, respectively, with 0.65-mm isotropic resolution.

The data are organized according to BIDS and include raw NIfTI images and linearly registered versions. The resource is available through Figshare. The original study was approved by the UNC School of Medicine Institutional Review Board, and all participants provided written informed consent. We use the paired T1w images from this resource for 3T-to-7T synthesis and process them with the unified pipeline described in Section 4.1 (Chen et al., 2023).

B.3.2 MULTIMODAL 3T–7T MRI DATASET

This dataset contains paired T1w, T2w, and resting-state fMRI scans from 20 healthy adults, including 10 men and 10 women aged 18–25 years. Each participant was scanned at 3T and 7T on the same day at the Beijing MRI Center for Brain Research. The 3T scans used a Siemens Prisma system with a 64-channel head coil. The 7T scans used a Siemens 7T system with a Nova 32-channel receive coil and single-channel transmit coil. The dataset also provides manual hippocampal-subfield annotations on the 7T T2w images.

The 3T and 7T T1w scans used MPRAGE sequences with 1.0-mm and 0.7-mm isotropic resolution, respectively. The data are organized according to BIDS and include MRIQC metrics and quality reports. The resource is freely available through Figshare+. All participants provided written consent for open sharing of anonymized data. We use the paired T1w images for synthesis and the 7T hippocampal annotations for the downstream evaluation (Li et al., 2024).

B.3.3 INTERNAL CLINICAL DATA

The collaborating hospital provided two internal clinical cohorts, each containing 25 paired T1w 3T–7T MRI cases, for a total of 50 cases. The data were de-identified and organized using random subject identifiers and the necessary imaging parameters; no directly identifying information was retained. The data are used for research only under the applicable data-use agreement and ethical requirements. The original clinical images are not released publicly.

B.3.4 USE IN THIS STUDY

All data were split by subject identity into training, validation, and test sets containing 56, 8, and 16 subjects, respectively, corresponding to 4,800, 675, and 1,200 slices. Slices from the same subject never appear in different splits. We apply the unified preprocessing pipeline described in Section 4.1. Registered images and quality reports distributed with the public resources are treated as dataset-provided materials and are not additional model inputs.

Table 6: Training and inference settings for the complete model.

Item	Setting
Input	2D single-channel, 320×256
Training / inference device	$2 \times$ RTX 3090 / $1 \times$ RTX 3090
Optimizer	AdamW
Base learning rate	1×10^{-4}
Backbone learning rate	2.5×10^{-5}
Detail and support branch learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Per-GPU batch size	1
Gradient accumulation steps	4
Effective batch size	8 slices
Maximum epochs	200
Early-stopping patience	30 epochs
Gradient clipping threshold	1.0
Training precision	Mixed precision; FP32 for band-related losses
Rollout sub-batch size	1 slice per GPU
Training time grid	$0, 0.2, 0.4, 0.6, 0.8$
Inference solver / steps	Euler / 5
Support-gate floor	0.2
Terminal low/mid/high weights	$1/2/4$
High-frequency RMS window	7×7
Intermediate-support RMS coefficient	0.5
Low-support high-frequency margin	0.1

B.4 Experimental Hyperparameters

Table 6 summarizes the training and inference settings of the complete model. The backbone is initialized from a base flow-matching checkpoint, while the detail and support-prediction branches are newly instantiated; all branches are then optimized jointly. The backbone uses a smaller learning rate so that it can adapt to the new branches while retaining the previously learned mapping. The maximum number of epochs is an upper bound, since training is stopped early according to the validation criterion.

During two-GPU training, each GPU uses a batch size of 1 and accumulates gradients for four steps, giving an effective batch size of eight slices per accumulation cycle. The checkpoint with the lowest validation total loss is selected for final evaluation; the test set is not used for checkpoint selection. The complete five-step rollout provides terminal supervision during training, whereas inference only performs the network updates. The weights of the local reconstruction and support losses are given in Appendix B.1.

C VISUALIZATION OF PREDICTED DETAIL SUPPORT

Figure 4 illustrates the spatial guidance described in Section 3.2. For each test slice, we show the source 3T image, synthesized 7T-like image, paired 7T reference, and predicted mid- and high-frequency support maps. We average the support predictions over the five Euler steps, denoting the resulting maps by $\bar{C}_M$ and $\bar{C}_H$. The maps display support before its transformation into soft gates. Support is predicted from source spatial and frequency features together with the current generation features; paired 7T images are shown only for reference and are not used during inference.

Within the brain, the maps exhibit spatially varying responses, with generally lower support in the high-frequency band than in the mid-frequency band for the displayed examples. This pattern is consistent with band-dependent modulation of additional refinement. Our qualitative interpretation focuses on the brain foreground. These maps describe learned guidance signals and do not establish the anatomical correctness of the synthesized details.

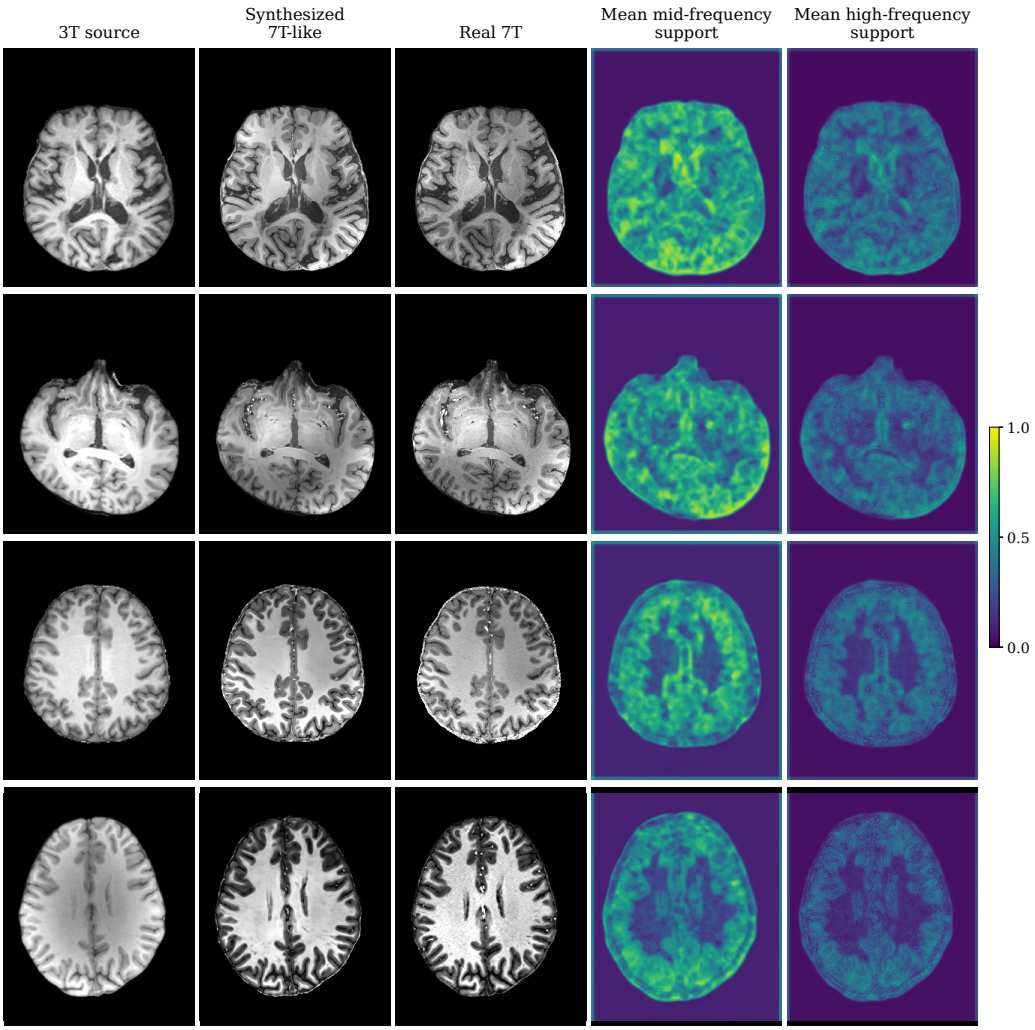


Figure 4: Visualization of predicted detail support on test slices. Each row shows, from left to right, the source 3T image, synthesized 7T-like image, paired real 7T image, and mid- and high-frequency support maps averaged over five inference steps. Both support maps use a shared color scale from 0 to 1, without per-image normalization.