

Prospective Dynamic 3D MRI Reconstruction via Latent-Space Motion Tracking from Single Measurement

Supplementary Material

Contents

1. Additional Analysis of Latent Vector	1
1.1. Visualization of Latent-Space Structure via t-SNE	1
1.2. More Interpolation Analysis of the Learned Latent Manifold	1
2. Inference Efficiency	1
3. Analysis of Tri-plane Feature	1
4. Dataset acquisition and pre-processing	2
4.1. XCAT Phantom	2
4.2. Inhouse Datasets	3
5. Implementation Details	4
6. Details of Baselines	4
6.1. Analytical Methods	4
6.2. Retrospective Reconstruction Methods	4
6.3. Prospective Reconstruction Methods	5
7. Additional Visual Results	6

1. Additional Analysis of Latent Vector

1.1. Visualization of Latent-Space Structure via t-SNE

To investigate whether the learned latent representation encodes physiologically meaningful motion patterns, we first discretize the respiratory motion signal (defined as the superior–inferior displacement of the liver’s center-of-mass) into 21 bins covering the full displacement range, as illustrated in Supplementary Fig. 1 (top). We then project the latent vectors corresponding to each time frame into a two-dimensional space using t-SNE. As shown in Supplementary Fig. 1 (bottom), the latent embeddings exhibit a smooth and ordered structure that aligns well with the motion-bin labels, indicating that the latent space forms a consistent low-dimensional manifold reflecting the underlying respiratory dynamics.

1.2. More Interpolation Analysis of the Learned Latent Manifold

To further examine the continuity and geometric structure of the learned latent manifold, we additionally perform spherical linear interpolation (slerp) between the latent vectors

corresponding to the end-inhale and end-exhale respiratory states. In contrast to linear interpolation in Euclidean space, slerp follows the geodesic on the hypersphere defined by latent vectors and therefore provides a more faithful interpolation path when the latent space exhibits nonlinear or curved geometry.

Let z_0 and z_1 denote the latent vectors at the end-inhale and end-exhale states, respectively. Given an interpolation parameter $t \in [0, 1]$, the slerp interpolation is defined as:

$$\text{slerp}(z_0, z_1; t) = \frac{\sin((1-t)\theta)}{\sin\theta} z_0 + \frac{\sin(t\theta)}{\sin\theta} z_1, \quad (1)$$

where the angular distance θ between the two latent vectors is

$$\theta = \arccos\left(\frac{z_0^\top z_1}{\|z_0\| \|z_1\|}\right). \quad (2)$$

Intermediate latent vectors are generated using uniform interpolation steps $t = 0.1, 0.2, \dots, 0.9$. The reconstructed images and their corresponding DVFs are shown in Supplementary Fig. 3. As t increases, the reconstructed anatomy evolves smoothly between inhale and exhale, and the inferred respiratory displacement varies in a physiologically coherent manner. The DVFs also exhibit a consistent and monotonic deformation progression, indicating that the latent space indeed supports a continuous and structured geodesic path aligned with true respiratory motion. These results further confirm that the learned manifold captures meaningful respiratory dynamics and generalizes well under nonlinear interpolation.

2. Inference Efficiency

Fig. 2 (L) reports the total runtime of each method using a method-specific number of iterations, which are selected to ensure convergence, as illustrated by the convergence curves in Fig. 2 (R). Tab. 1 reports the per-iteration runtime (ms) of different methods, evaluated on NVIDIA A100 GPU, where each method is executed 100 times and the mean and standard deviation are reported. As shown, our proposed PDMR achieves a more favorable balance between reconstruction quality and computational cost compared to other methods.

3. Analysis of Tri-plane Feature

For tri-plane features $F \in \mathbb{R}^{3 \times C \times H \times W}$, we apply spatial PCA to each plane independently. Taking the XY plane as an example, We treat each pixel as one sample with a

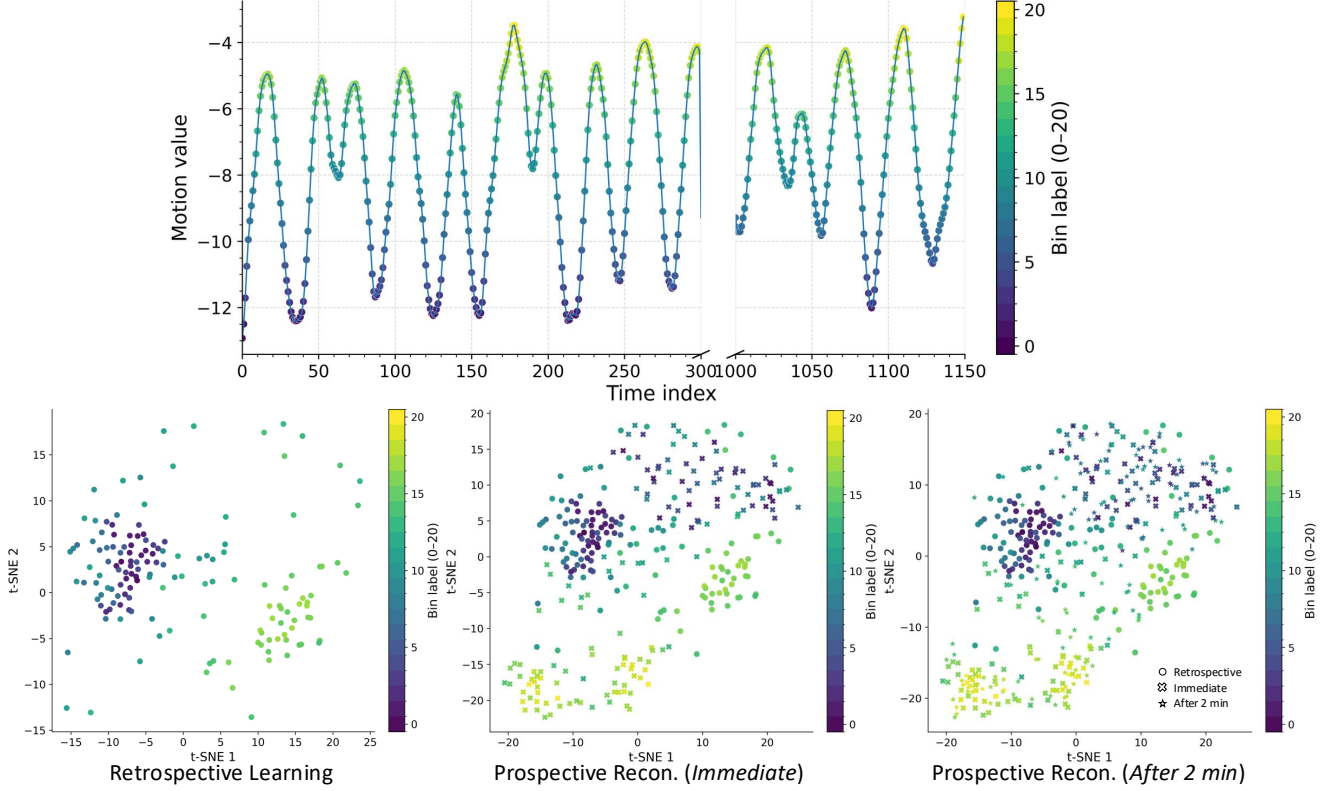


Figure 1. t-SNE analysis of learned latent vectors. Top: Respiratory motion signal (superior–inferior displacement of the liver center-of-mass) discretized into 21 bins, shown with color-coded bin labels. Bottom: t-SNE embeddings of latent vectors obtained from (left) retrospective learning, (middle) prospective reconstruction (immediate), and (right) prospective reconstruction after a 2-minute gap. Colors indicate the corresponding motion bin (0–20). Circles denote retrospective points, crosses denote prospective–immediate points, and stars denote prospective–after-2-min points.

NUFFT	GRASP	TDDIP	SPINER	MOTUS	Ours
Time 7.39 ± 0.00	75.82 ± 56.10	370.09 ± 6.17	307.44 ± 39.19	83.35 ± 8.89	216.09 ± 2.90

Table 1. Per-iteration runtime of different methods (ms).

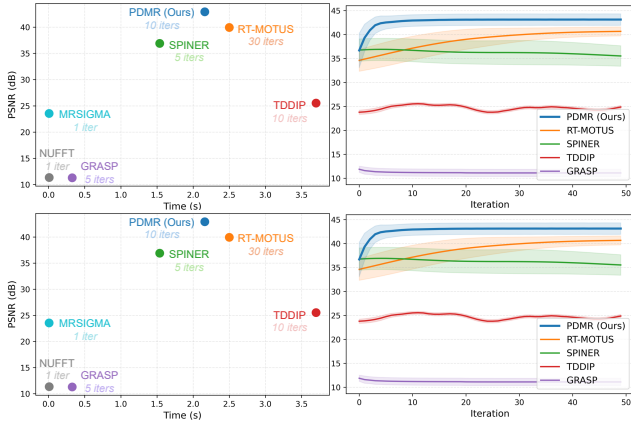


Figure 2. Performance and inference efficiency comparisons.

C -dimensional feature vector, forming $\mathbf{F}_{xy} \in \mathbb{R}^{(H \cdot W) \times C}$. After zero-centering \mathbf{F}_{xy} , we compute the economy SVD

$\mathbf{F}_{xy} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$. For visualization, we reshape the PCA scores $\mathbf{U}\mathbf{S}$ of the first three components to $H \times W$ and stack them as R/G/B=PC1/PC2/PC3 to form a single pseudo-RGB map per plane. Maps are min–max normalized per component for display only. We perform this procedure separately for the End-inhale and End-exhale features under identical settings, enabling direct qualitative comparison of dominant spatial modes across respiratory states. The visualization can be seen in Fig. 4.

4. Dataset acquisition and pre-processing

4.1. XCAT Phantom

XCAT, short for the Extended Cardiac Torso model [12], is a highly detailed, anatomically accurate computational phantom widely used in medical imaging research and simulation. Developed to provide realistic representations of human anatomy and physiological motion, XCAT integrates both cardiac and respiratory dynamics, enable to simulate complex scenarios that closely resemble real clinical conditions.

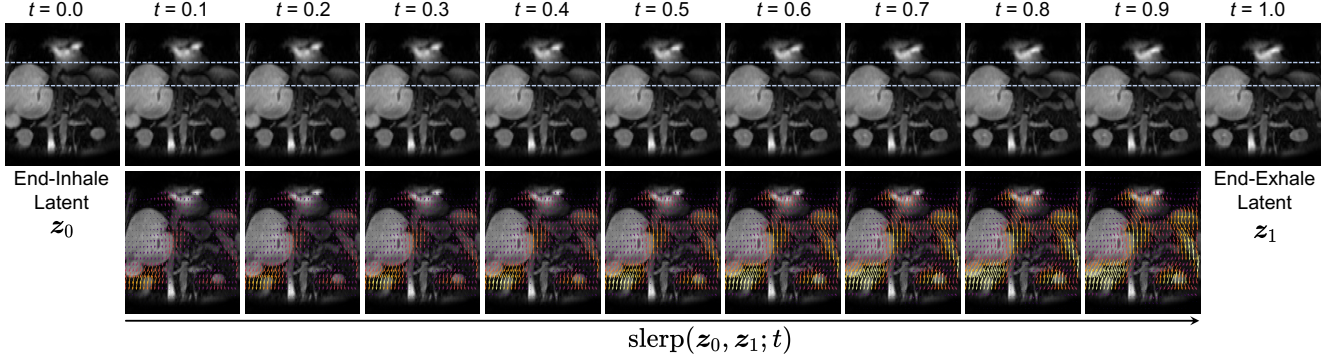


Figure 3. Reconstructed images (top) and DVFs (bottom) generated by spherical linear interpolation between the latent vectors z_0 and z_1 for $t=0.0-1.0$. The horizontal dashed line indicates the inferred respiratory displacement, which varies smoothly across interpolation steps. The smooth anatomical transition and consistent DVF progression indicate that the learned latent space encodes a continuous and physiologically meaningful respiratory manifold.

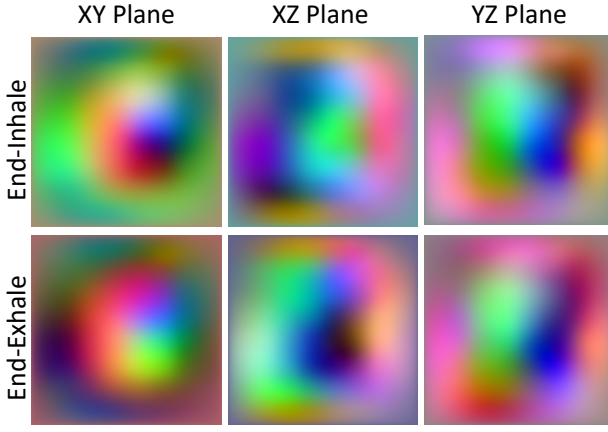


Figure 4. Spatial PCA pseudo-RGB maps ($R/G/B = PC1/PC2/PC3$) of tri-plane features at End-Inhale (top) and End-Exhale (bottom). Columns show XY / XZ / YZ planes; color shifts highlight state-dependent dominant spatial modes.

We manually selected the abdominal section of the XCAT phantom with a spatial resolution of $1.6 \times 1.6 \times 2.0 \text{ mm}^3$ and a matrix size of $224 \times 224 \times 96$. For respiratory motion, we defined a 3.96s breathing cycle and generated 500 frames at a temporal resolution of 170 ms per frame. After generating the XCAT phantom, we assigned T_1 , T_2 , and M_0 values to the corresponding anatomical regions based on values reported in the literature. Using these tissue parameters, we then simulated a spoiled gradient echo acquisition following the same sequence settings as the in-house data. The resulting fully sampled simulated images were used as the ground-truth data for subsequent experiments.

4.2. Inhouse Datasets

Under institutional review board approval, 6 DCE-MRI examinations of 6 patients were performed as part of a pilot study of individualized adaptive radiation therapy for hepatocellular carcinoma. A 3-T MRI scanner (Magnetom Skyra, Siemens Healthineers, Erlangen, Germany) was used. As part of the scan protocol, a 10-min DCE-MRI scan was performed using a work-in-progress golden-angle stack-of-stars spoiled gradient echo sequence [1, 4] with fat suppression. A 20 ml (0.5 M) of Gd-BOPTA (MultiHance, Bracco Diagnostics, Monroe, NJ) was administered 30 s after the start of scanning. For reception, an 18-channel flexible surface coil (Body Matrix) was used in combination with 2–5 elements of the posterior coils built into the scanner table (Spine Matrix). Prior to scanning, a calibration scan was used to determine a receiver-coil noise whitening transform [11] as well as a set of coil sensitivities [2]. After calibration, subjects were scanned with 3500 radial through-center spokes. Sequence parameters are listed in Table 2. The sequence collected 46 Cartesian partitions in the SI direction, covering three-fourths of k-space with 384 samples per line. The central partition was used to determine a gradient-delay correction by comparing lines acquired in opposite directions for the latter half of the number of acquired spokes. The correction shifted acquired spokes by modulating their Fourier transform with a complex wave. After delay correction, the missing one-fourth of k-space was synthesized using a partial-Fourier projectiononto-convex-sets technique to produce 58 partitions. The noise whitening transform determined from the calibration scan was then used to transform the coil signals into synthetic signals with independent and identically distributed noise.

The reference motion signal was derived from an image time series with high temporal but lower spatial resolution.

Sequence parameter	
Echo time	1.14–1.21 ms
Repetition time	2.72–4.51 ms
Flip angle	10°–14°
Image matrix size	192 × 192
Number of slices	64
Number of partitions	46
Number of radial spokes	3500
In-plane voxel size	2–2.45 mm
Slice thickness	3–4 mm

Table 2. In-house Data Sequence Parameters.

The resulting 3500 images were rigidly aligned with respect to a reference image in an arbitrary breathing state using a robust region-limited rigid-body image registration algorithm [8] with translation but no rotation. The reference image was selected among the VEN images by a physician. The superior–inferior (SI) translation of the center of mass of the liver was extracted from each of the 3500 transforms produced by the registration and used as a one-dimensional motion signal. The motion signal for the subjects can be seen in Fig. 5.

5. Implementation Details

We implemented our generator \mathcal{G}_ψ on top of the official PyTorch implementation of StyleGAN2 [9] (<https://github.com/NVlabs/stylegan3>). The mapping network consists of 6 fully connected layers, while the synthesis network has a channel base of 32768 and a maximum channel width of 512. We disable FP16 layers and no output clamping is applied. The tri-plane decoder is implemented as a lightweight MLP composed of 5 layers with a hidden width of 128.

6. Details of Baselines

We compare our PDMR model with six representative methods. Here, we provide the implementation details of these baselines to improve the reproducibility of this work.

6.1. Analytical Methods

Both NUFFT [6] and GRASP [5] are analytical reconstruction methods and therefore require no retrospective training. Accordingly, we directly apply them to perform prospective reconstruction on each single measurement.

NUFFT Non-uniform fast Fourier transform (NUFFT) [6] is an analytical reconstruction algorithm designed for MRI with non-uniform sampling patterns, such as golden-angle radial sampling. It first uses an interpolation algorithm (e.g., linear) to generate uniform

k-space data, followed by applying the IFFT operator to reconstruct the final MR images. We implement it using the function `KbNufftAdjoint` from the Python library `torchkbnufft` (<https://github.com/mmuckley/torchkbnufft>).

GRASP GRASP [5] is a compressed sensing framework for golden-angle radial sampling that jointly reconstructs images with temporal sparsity constraints. We use the official MATLAB implementation provided by the authors (<https://cai2r.net/resources/grasp-matlab-code/>).

6.2. Retrospective Reconstruction Methods

For retrospective methods, we first train each model retrospectively on spokes 0–150. During prospective reconstruction, the model is then fine-tuned using only the current measurement. The implementation details for each baseline are provided below. While we aim to ensure the fairest possible comparison between retrospective approaches, achieving perfect fairness is inherently difficult. This also highlights the practical challenges and limitations faced by existing retrospective methods when transferred to a prospective reconstruction setting.

TDDIP TDDIP [14] models the temporal evolution of a dynamic sequence by learning a one-dimensional manifold, parameterized by time, that maps to the corresponding dynamic images via a CNN-based generative network. For retrospective learning, we follow the original implementation and adopt spoke-shared measurements with a window size of 10, as we found that training with single-spoke resolution leads to noticeably degraded performance. For prospective reconstruction, the pretrained CNN generator is frozen, and we optimize only the one-dimensional latent parameter using the current single measurement. We note that the comparison may not be entirely fair: unlike motion-compensated methods, TDDIP does not utilize a template image as an explicit prior, which introduces an inherent mismatch when transferring the method to a prospective reconstruction setting. We use the official implementation provided by the authors (<https://github.com/jaejun-yoo/TDDIP/>).

SPINER SPINER [3] is a motion-compensated reconstruction method that models both the template image and the deformation vector fields (DVF) using implicit neural representations (INRs). To ensure a fair comparison, we do not train an INR to learn the template image; instead, we use the pre-scan reconstructed image as the fixed template. For retrospective learning, the time coordinates corresponding to spokes 0–150 are linearly scaled to the interval $(-0.5, 0.5)$ and used as inputs to the INR to learn the DVFs. For *immediate* prospective reconstruction, we shift the temporal input to $(0.5, 1.5)$, and for the *after-2min* scenario, we

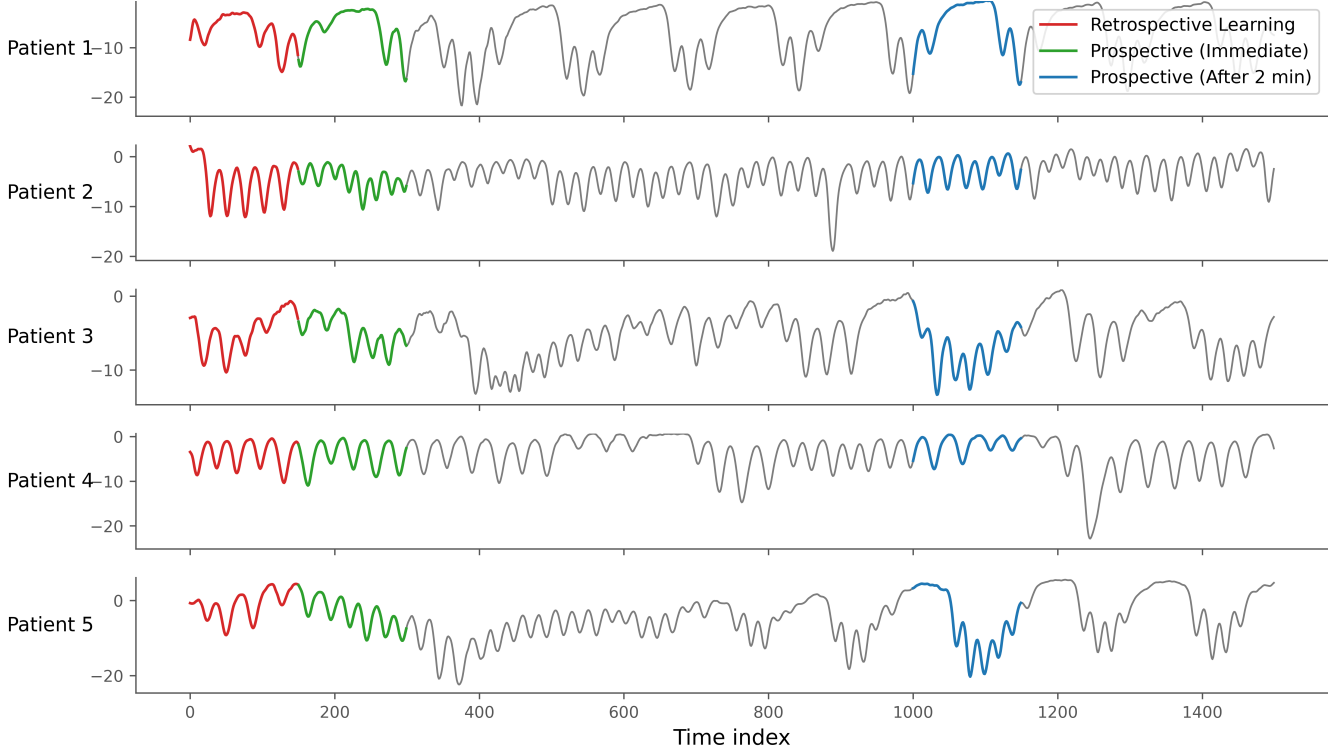


Figure 5. For each patient, the superior–inferior respiratory displacement is shown over time, with segments corresponding to different reconstruction settings highlighted: red indicates the portion used for retrospective learning, green denotes the prospective (immediate) reconstruction period, and blue marks the prospective reconstruction performed after a 2-minute gap.

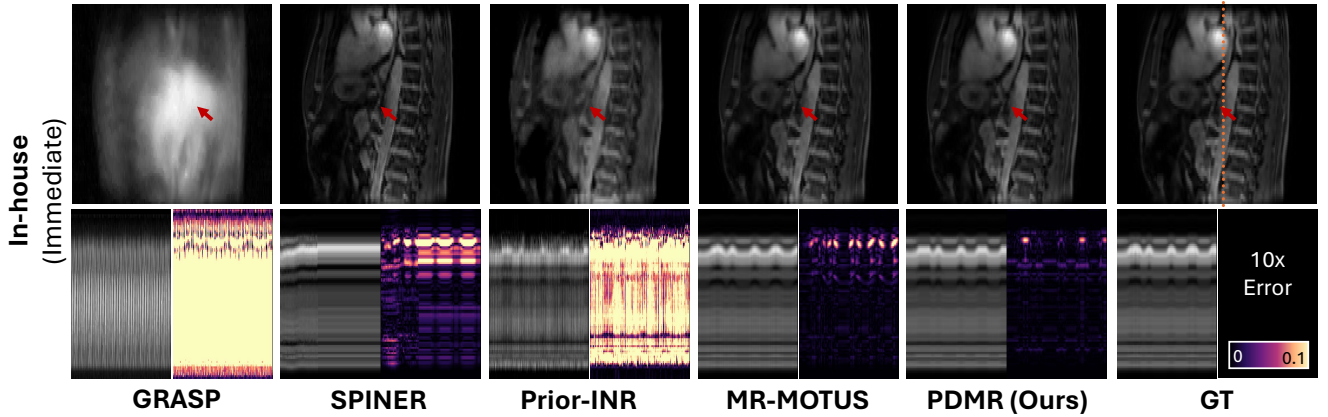


Figure 6. Additional sagittal-view prospective reconstruction results.

shift it to (3.333, 4.333). In both cases, the pretrained INR is further fine-tuned using only the current single measurement. It is worth noting that, in real prospective reconstruction, the true temporal coordinate is not directly observable. This reflects a fundamental limitation of INR-based methods that explicitly rely on time coordinates as inputs when applied to prospective reconstruction settings.

6.3. Prospective Reconstruction Methods

Prior-INR Prior-INR [10] first fits the inhale state ($t = 0$) and the exhale state ($t = 1$) using separate implicit neural representations (INRs). A set of intermediate images is then generated by interpolating and extrapolating across the temporal axis, producing a discrete set of thirty images spanning $t = -1, -0.9, \dots, 1.9, 2$. During prospective reconstruction, we identify the image in this discrete set

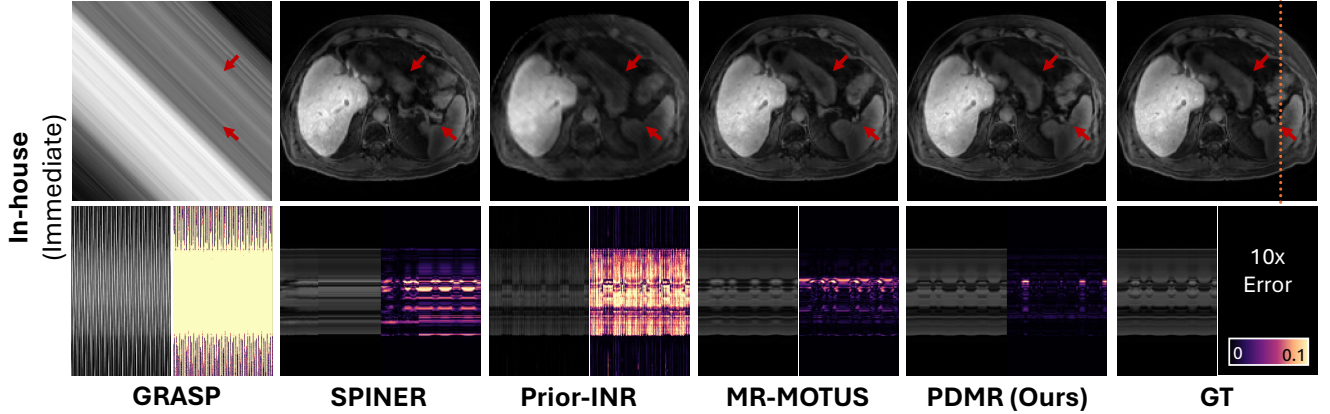


Figure 7. Additional axial-view prospective reconstruction results.

whose forward projection best matches the current single measurement, and use that image as a prior. The network is then fine-tuned using the current measurement. However, because the prior images are sampled from a discrete set rather than a continuous manifold, the resulting reconstruction often exhibits temporal discontinuities, reflecting a fundamental limitation of this prior-based discrete INR strategy. It is worth noting that Prior-INR requires substantially more prior information than other baselines, since it relies on both inhale and exhale images to construct the discrete prior set.

MR-MOTUS MR-MOTUS [7, 13] models the deformation vector fields (DVs) as a low-rank combination of spatial and temporal components, where each component is parameterized using B-spline bases. Following the original configuration, we adopt 24 spatial B-splines and 5 temporal B-splines. However, we observed that using the original rank of 1 leads to suboptimal performance in our setting; therefore, we increase the rank to 2 for a fairer comparison. MR-MOTUS is fundamentally limited by the expressive capacity of its linear motion manifold, and its hand-crafted model design introduces numerous hyperparameters that require manual tuning.

7. Additional Visual Results

Fig. 6 and Fig. 7 provide additional qualitative results in the sagittal and axial views, respectively. These complementary visualizations further demonstrate the anatomical consistency and motion-robust reconstruction quality achieved by our method across different imaging planes.

We provide **reconstruction videos**, including sagittal view and coronal view, to better illustrate the dynamic changes (please refer to the attached videos: *video-coronal.mp4* and *video-sagittal.mp4*).

References

- [1] Kai Tobias Block, Hersh Chandarana, Sarah Milla, Mary Bruno, Tom Mulholland, Girish Fatterpekar, Mari Hagiwara, Robert Grimm, Christian Geppert, Berthold Kiefer, et al. Towards routine clinical use of radial stack-of-stars 3d gradient-echo sequences for reducing motion sensitivity. *Investigative Magnetic Resonance Imaging*, 18(2):87–106, 2014. 3
- [2] William W Brey and Ponnada A Narayana. Correction for intensity falloff in surface coil magnetic resonance imaging. *Medical physics*, 15(2):241–245, 1988. 3
- [3] Lixuan Chen, James M Balter, Liye Shen, and Jeong Joon Park. Single-spoke motion-compensated dynamic 3d mri reconstruction via neural representation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 513–522. Springer, 2025. 4
- [4] Li Feng. Golden-angle radial mri: basics, advances, and applications. *Journal of Magnetic Resonance Imaging*, 56(1): 45–62, 2022. 3
- [5] Li Feng, Qiuting Wen, Chenchuan Huang, Angela Tong, Fang Liu, and Hersh Chandarana. Grasp-pro: improving grasp dce-mri through self-calibrating subspace-modeling and contrast phase automation. *Magnetic resonance in medicine*, 83(1):94–108, 2020. 4
- [6] Jeffrey A Fessler and Bradley P Sutton. Nonuniform fast fourier transforms using min-max interpolation. *IEEE transactions on signal processing*, 51(2):560–574, 2003. 4
- [7] Niek RF Huttinga, Tom Bruijnen, Cornelis AT Van Den Berg, and Alessandro Sbrizzi. Real-time non-rigid 3d respiratory motion estimation for mr-guided radiotherapy using mr-motus. *IEEE Transactions on Medical Imaging*, 41(2):332–346, 2021. 6
- [8] Adam Johansson, James Balter, Mary Feng, and Yue Cao. An overdetermined system of transform equations in support of robust dce-mri registration with outlier rejection. *Tomography*, 2(3):188, 2016. 4
- [9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of*

the *IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. [4](#)

- [10] Lianli Liu, Liyue Shen, Adam Johansson, James M Balter, Yue Cao, Lucas Vitzthum, and Lei Xing. Volumetric mri with sparse sampling for mr-guided 3d motion tracking via sparse prior-augmented implicit neural representation learning. *Medical physics*, 51(4):2526–2537, 2024. [5](#)
- [11] N Martini, MF Santarelli, G Giovannetti, M Milanesi, D De Marchi, V Positano, and Luigi Landini. Noise correlations and snr in phased-array mrs. *NMR in Biomedicine: An International Journal Devoted to the Development and Application of Magnetic Resonance In vivo*, 23(1):66–73, 2010. [3](#)
- [12] W Paul Segars, G Sturgeon, S Mendonca, Jason Grimes, and Benjamin MW Tsui. 4d xcat phantom for multimodality imaging research. *Medical physics*, 37(9):4902–4915, 2010. [2](#)
- [13] Hua-Chieh Shao, Xiaoxue Qian, Guoping Xu, Can Wu, Ricardo Otazo, Jie Deng, and You Zhang. A dynamic reconstruction and motion estimation framework for cardiorespiratory motion-resolved real-time volumetric mr imaging (dreme-mr). *arXiv preprint arXiv:2503.21014*, 2025. [6](#)
- [14] Jaejun Yoo, Kyong Hwan Jin, Harshit Gupta, Jerome Yerly, Matthias Stuber, and Michael Unser. Time-dependent deep image prior for dynamic mri. *IEEE Transactions on Medical Imaging*, 40(12):3337–3348, 2021. [4](#)