Prior-Adapted Progressive Time-Resolved CBCT Reconstruction Using a Dynamic Reconstruction and Motion Estimation Method

Ruizhi Zuo¹, Hua-Chieh Shao¹, You Zhang¹

¹The Medical Artificial Intelligence and Automation (MAIA) Laboratory
Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390,
USA

Corresponding address:

You Zhang
Department of Radiation Oncology
University of Texas Southwestern Medical Center
2280 Inwood Road
Dallas, TX 75390
Email: You.Zhang@UTSouthwestern.edu

Tel: (214) 645-2699

Abstract

Background:

Cone-beam CT (CBCT) provides on-board volumetric anatomy for image guidance and treatment adaptation in radiotherapy. To compensate for respiration-induced anatomical motion, time-resolved CBCT is highly desired to capture the spatiotemporal anatomical variations but faces challenges in accuracy and efficiency due to substantial optimization needed in image reconstruction and motion modeling.

Purpose:

We proposed a fast time-resolved CBCT reconstruction framework, based on a dynamic reconstruction and motion estimation method with new reconstructions initialized and conditioned on prior reconstructions in an adaptive fashion (DREME-adapt).

Materials and Methods:

DREME-adapt reconstructs a time-resolved CBCT sequence from a fractional standard CBCT scan while simultaneously generating a machine learning-based motion model that allows single-projection-driven intra-treatment CBCT estimation and motion tracking. Via DREME-adapt, a virtual fraction is generated from a pre-treatment 4D-CT set of each patient for a clean, 'cold-start' reconstruction. For subsequent fractions of the same patient, DREME-adapt uses pre-derived motion models and reference CBCTs as initializations to drive a 'warm-start' reconstruction, based on a lower-cost refining strategy. Three strategies: DREME-cs which drops the 'warm-start' component, DREME-adapt-vfx which uses a fixed initialization (virtual fraction's reconstruction results), and DREME-adapt-pro which initialize reconstructions through a progressive daisy chain scheme (virtual fraction for fraction 1, fraction 1 for

fraction 2, and so on), were evaluated on a digital phantom study (7 motion/anatomical scenarios) and a patient study (7 patients).

Results:

DREME-adapt allows fast and accurate time-resolved CBCT reconstruction. For the XCAT simulation study, DREME-adapt-pro achieves image reconstruction relative error of 0.14±0.01 and tumor center-of-mass tracking error of 0.92±0.62 mm (mean±s.d.), compared to 0.15±0.01 and 1.06±0.73 mm for DREME-adapt-vfx, and 0.18±0.01 and 1.96±1.35 mm for DREME-cs. For the real-time motion inference test dataset of the patient study, DREME-adapt-pro localizes moving lung landmarks to a mean±s.d. error of 2.21±1.79 mm. In comparison, the corresponding values for DREME-adapt-vfx and DREME-cs were 2.53±1.93 mm and 3.22±2.88 mm, respectively. The DREME-adapt-pro training takes 11 minutes, only 15% of the original DREME algorithm.

Conclusions:

With high efficiency and accuracy, DREME-adapt-pro allows on-board time-resolved CBCT reconstruction and enhances the clinical adoption potential of the DREME framework.

Keywords: Time-resolved dynamic CBCT, Image reconstruction, Motion estimation, Motion model, Fine-tuning.

1. Introduction

Cone-beam computed tomography (CBCT) is widely used in clinical practice, offering on-board high spatial resolution volumetric imaging guidance for treatment setup, dose verification, and adaptive radiotherapy^{1,2}. A typical CBCT scan takes about one minute for a full 360° acquisition. However, patient anatomical motion, mainly due to respiration with a cycle of 3-5 seconds, can introduce motion artifacts and blurring in the reconstructed images^{3,4}. To address this concern and resolve the underlying motion, four-dimensional (4D) CBCT was developed as the current clinical solution. 4D-CBCT organizes the projections into predefined respiratory phases and assembles semi-static CBCTs reconstructed from each bin into a sequence of respiration-resolved 3D CBCT images. However, achieving high 4D-CBCT image quality is challenging due to streak artifacts and blurry tissue boundaries from phase-sorting-induced undersampling⁵. Iterative methods, especially those combined with compressive sensing ⁶, were proposed to suppress the undersampling artifacts by minimizing differences between target images and reconstructed images in a sparse space. However, their efficacy is limited, especially when sorted projections are highly limited. Another approach is to build a motion model from prior images, such as a patient-specific 4DCT^{7,8}, to achieve dimension reduction and enable few projection-based, deformation-driven image reconstruction. However, the deformation-based approach faces the challenge of the intrinsic differences in motion patterns, anatomy, and intensities between the prior and target images⁹. More importantly, the motion sorting/binning of conventional 4D-CBCT assumes that the patient motion is regular and periodic, which is generally false and limits its ability to capture time-resolved irregular motion¹⁰. To resolve both regular and irregular motion to guide radiotherapy delivery, time-resolved CBCT imaging, by which each X-ray projection is reconstructed into its own CBCT image, is highly desired. Time-resolved dynamic CBCT imaging provides the highest spatiotemporal resolution that can be achieved with an X-ray scanner, but is difficult for conventional reconstruction methods due to extreme undersampling.

In the past few years, deep learning (DL)-based approaches have become popular in CBCT reconstruction because of the low inference latency and improved accuracy¹¹⁻¹⁵. DL solutions have demonstrated the

feasibility of 3D image reconstruction from highly sparse 2D X-ray projections¹⁶⁻¹⁸. Ying et al.¹² proposed reconstructing a CBCT from two orthogonal X-rays using a generative adversarial network (GAN). The generator consists of 2D encoders, a fusion module, and a 3D decoder to generate the 3D volume. Although the GAN-based model can generate high-quality images, the fine details and structure may not be reliable due to the synthetic nature of the GAN models. Shen et al.¹⁹ introduced a network integrating geometric priors of the imaging system to bridge 2D and 3D image space to assist CBCT image reconstruction from ultra-sparse projections. However, DL methods based on the sparse measurements did not provide sufficient information for precise reconstruction due to the ill-posed nature of the inverse problem, resulting in poor generalization across diverse datasets. Additionally, the lack of 'ground truth' clinical CBCT datasets may further limit the potential of such supervised learning, which usually requires a large training dataset.

Implicit neural representation (INR)²⁰ learning has emerged as a novel machine learning technique to map spatial coordinates to corresponding physical quantities (for instance, CBCT image intensities), serving a universal function approximator that enables high-resolution and memory-efficient representations of volumetric images. Directly representing image volumes as neural networks (rather than gridded voxels). INR-based strategy allows the use of deep learning frameworks to efficiently and effectively optimize the image-representing networks for iterative, limited-sample-based image reconstruction. Based on INR, the simultaneous Spatial and Temporal Implicit Neural Representation (STINR) framework²¹ decouples the time-resolved dynamic CBCT reconstruction into solving a spatial INR to represent a reference static CBCT image, and temporal INRs to learn intra-scan, time-resolved dynamic motion from each X-ray projection. The solved dynamic motion, after being applied to the reference CBCT, yields projection-specific dynamic CBCT images. Such a motion-compensated reconstruction strategy uses all available projections via a 'oneshot' learning and self-supervised framework to simultaneously optimize the spatial and temporal INRs. It effectively addresses the undersampling challenge to reconstruct a high-quality dynamic CBCT set without the need of population-based large-scale training. Based on STINR, Shao et al. further improves the reconstruction strategy by removing its use of a fixed prior motion model, yielding a prior-model-free STINR (PMF-STINR) technique. PMF-STINR learns an on-the-fly motion model during the reconstruction, the efficacy of which has been validated on multi-institutional real patient scans²². Recently, on the basis of PMF-STINR, Shao et al. further proposed a joint dynamic reconstruction and motion estimation (DREME) framework for real-time CBCT imaging and motion tracking²³. In addition to reconstructing a dynamic CBCT set and deriving a motion model from a pre-treatment 3D CBCT scan, DREME simultaneously learns a convolutional neural network (CNN)-based motion encoder, which can infer motion model coefficients from arbitrarily angled, singular X-ray projections to derive real-time deformable motion. The dual-task learning system of DREME enables real-time 3D deformable motion and CBCT inference from later-acquired intra-treatment X-ray projections for continuous motion monitoring and tracking. While achieving high accuracy in dynamic reconstruction and real-time motion tracking, the training time for DREME is approximately 3-4 hours (based on CBCTs of 200×200×100 voxels) on an Nvidia V100 GPU card (~73 mins on Nvidia RTX 4090). The time-consuming nature of DREME, driven by prolonged iterations for from-scratch 'one-shot' motion modeling and image reconstruction, hinders its clinical practicality when on-the-fly reconstruction is needed at each treatment fraction.

To overcome the efficiency limitations of DREME reconstruction, in this study we developed a prior-adapted reconstruction strategy based on the DREME framework (DREME-adapt). By DREME-adapt, patient-specific pre-treatment 4D-CT images are used to simulate motion-involved cone-beam projections to mimic a 'virtual' treatment fraction, and the projections are trained by the original DREME framework to yield a virtual-fraction DREME model. DREME-adapt then uses the reference CBCT and the motion model from the virtual-fraction DREME reconstruction as 'warm-start' initializations, to jump-start and speed-up the dynamic CBCT reconstructions at real treatment fractions. Under the framework of DREME-

adapt, we investigated two fine-tuning strategies: 1. DREME-adapt-vfx, which always uses the virtual-fraction DREME model to warm-start the CBCT reconstruction at each treatment fraction; 2. DREME-adapt-pro, which initializes the 1st treatment fraction with the virtual-fraction DREME model. The DREME model derived from the 1st treatment fraction is then applied to initialize the reconstruction of the 2nd treatment fraction, forming a progressive daisy chain. In addition, we also evaluated the original DREME framework without the 'warm-start' component for comparison (DREME-cs, 'cold-start'). All strategies were evaluated with both a digital phantom simulation study and a lung patient dataset study. Compared with the original DREME reconstruction, DREME-adapt-pro reduces the reconstruction time by 85% (~11 mins on RTX 4090) while preserving high reconstruction accuracy. With enhanced efficiency and preserved accuracy, DREME-adapt-pro demonstrates its potential for clinical translation in dynamic and real-time CBCT imaging.

2. Materials and Methods

DREME-adapt reconstructs a time-resolved CBCT sequence from a fractional standard CBCT scan while simultaneously generating a machine-learning-based motion model that allows single-projection-driven intra-treatment CBCT estimation and motion tracking. Via DREME-adapt, we use the pre-treatment 4D-CT scan of a patient to simulate motion-involved cone-beam projections for a clean 'cold-start' reconstruction. For subsequent real treatment fractions of the same patient, DREME-adapt uses pre-derived motion models and reference CBCT as initializations to drive 'warm-start' reconstructions, based on a lower-cost refining strategy.

2.1 Dynamic and real-time CBCT reconstruction by the original DREME framework

In general, the original DREME framework reconstructs a dynamic sequence of 3D CBCT volumes from a pre-treatment projection set p. Due to the high spatiotemporal correlations within a dynamic CBCT set, the dynamic anatomy could be decomposed as a static reference anatomy $I_{ref}(x)$ and a time-varying, projection-dependent deformable vector field (DVF) set d(x, p) on the reference anatomy, i.e.,

$$I(x,p) = I_{ref}(x + d(x,p)), \tag{1}$$

where x denotes the voxel coordinates and $p \in p$. Since solving time-dependent DVFs remains an ill-posed spatiotemporal inverse problem, we further decomposed DVFs as a summation of products between projection-specific temporal components $(w_i(p))$ and projection-agnostic spatial components $(e_i(x))$, rendering a motion model:

$$d(x,p) = \sum_{i=1}^{3} w_i(p) \times e_i(x).$$
 (2)

By DREME, the reference anatomy $I_{ref}(x)$ is represented by a spatial INR that applies mapping $x \mapsto I_{ref}(x)$ from a voxel coordinate x to the corresponding attenuation coefficient $I_{ref}(x)$. $w_i(p)$ is represented by temporal motion coefficients derived by a CNN-based motion encoder, and $e_i(x)$, represented by a B-spline interpolant, captures the spatial motion variations that compose a basis set (motion basis components, MBCs). Three levels (i = 1,2,3) are used here since they were demonstrated sufficient for representing respiratory motion²⁴. The spatial INR, the CNN-based motion encoder, and the

MBCs are all learned on-the-fly directly from the raw cone-beam projections of a pre-treatment CBCT scan, via a 'one-shot' and self-supervised learning scheme. The workflow and the network architecture of the DREME framework is summarized in Fig. 1. DREME has two learning tasks. The first learning task (Fig. 1, upper panel) is dynamic CBCT reconstruction, where acquired cone-beam projections serve as input to generate a dynamic CBCT sequence. The reconstruction loss is defined based on the acquired projections, along with multiple regularization losses on $I_{ref}(x)$, MBC components $e_i(x)$, and their scores $w_i(p)$ (Eqs. 1&2). The second learning task (Fig. 1, middle panel) is to further train the CNN motion encoder to be robust to different motion scenarios (especially those not contained in the acquired projections) for accurate real-time motion estimation. In the on-board real-time CBCT inference stage (Fig. 1, lower panel), the intratreatment cone-beam projection p is fed into the network to infer the MBC scores and the corresponding real-time CBCT and DVFs. If I_{ref} here is replaced by the contour of a tracking target, DREME can be used for real-time target localization.

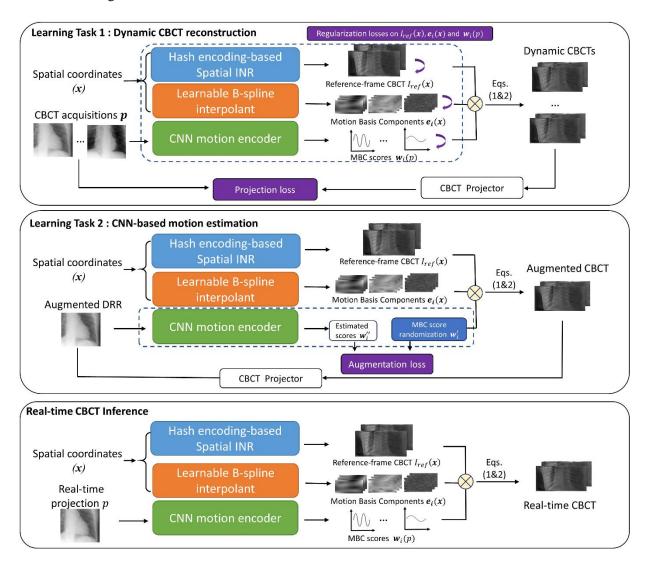


Figure 1. The DREME framework for dynamic CBCT reconstruction and real-time motion estimation. In learning task 1 (*upper panel*), a spatial implicit neural representation (INR) network represents the reference CBCT $I_{ref}(x)$. The learnable B-spline interpolant and the CNN motion encoder represent the spatial $e_i(x)$

and temporal components $\mathbf{w}_i(p)$ of the deformation vector fields $\mathbf{d}(\mathbf{x},p)$, respectively. From the three components, time-resolved dynamic CBCTs are derived using Eqs. 1 and 2. The reconstruction is driven by the projection loss between the CBCT acquisition \mathbf{p} and digitally reconstructed radiographs (DRRs) projected from the derived dynamic CBCT sequence. Different regularization losses are incorporated into the training to address the ill-posed nature of the spatiotemporal inverse problem. The learning parameters enclosed within the blue dashed box are subject to update during the training. In learning task 2 (middle panel), deformable motion augmentation is introduced by randomizing the temporal coefficients of the motion model (\mathbf{w}_i') to sample random motion. The projection angle augmentation is also implemented to simulate DRRs at random projection angles, to make the CNN motion encoder robust to arbitrary projection directions. The real-time CBCT inference stage (lower panel) takes an on-board cone-beam projection p from an arbitrary angle as input, to estimate the MBC scores and the corresponding real-time CBCT and motion field. If the reference CBCT is replaced by the contour of a tracking target, DREME can be used for real-time target localization.

2.2 Loss function design and training strategy of DREME

DREME uses a progressive training strategy with increasing training complexity to improve learning efficiency and avoid overfitting. It consists of a low-resolution stage (stage I) with an isotropic $4\times4\times4$ mm³ reconstruction resolution and a high-resolution stage (stage II) with a $2\times2\times2$ mm³ resolution, each containing multiple substages, with distinct total loss functions combining differently weighted loss components.

During the first training substage (I-a), the spatial INR model is initialized using a motion-blurred, approximate CBCT $I_{app}(x)$, reconstructed using all available projections by the Feldkamp-Davis-Kress method²⁵. The image-domain similarity loss L_{sim}^{im} between $I_{ref}(x)$, which is voxelized from the spatial INR, and $I_{app}(x)$ is calculated as:

$$L_{sim}^{im} = \frac{1}{N_{voxel}} \sum_{l=1}^{N_{voxel}} |I_{ref}(x) - I_{app}(x)|^2,$$
 (3)

where N_{voxel} denotes the number of voxels in the CBCT. In the next substage I-b, the spatial INR is then optimized with a projection domain similarity loss L_{sim}^{prj} between the digitally reconstructed radiographs (DRRs) generated from the reference CBCT $I_{ref}(x)$ by the projector \mathcal{P} and the corresponding cone-beam projections \mathbf{p} :

$$L_{sim}^{prj}(\boldsymbol{I}_{ref}(\boldsymbol{x})) = \frac{1}{N_{batch} N_{pixel}} \sum_{N_{batch} N_{pixel}} \left| \mathcal{P}[\boldsymbol{I}_{ref}(\boldsymbol{x})] - \boldsymbol{p} \right|^{2}, \tag{4}$$

where N_{batch} and N_{pixel} are the number of projection samples in a training batch and the number of pixels in a projection, respectively. In addition to the similarity loss, DREME incorporates the total variation (TV) ²⁶ regularization into the CBCT reconstruction to suppress high-frequency image noise while preserving anatomical edges:

$$L_{TV} = \frac{1}{N_{voxel}} \sum_{N_{voxel}} |\nabla I_{ref}(\mathbf{x})|, \qquad (5)$$

where ∇ denotes the gradient operator.

In the substage I-c, DREME added the motion model into training, using different regularization losses on the MBCs $e_i(x)$ and MBC scores $w_i(p)$. For the MBCs $(e_i(x))$, it enforced their ortho-normality as:

$$L_{MBC} = \frac{1}{9} \sum_{k=x,y,z} \sum_{i=1}^{3} \left(\left| \left\| e_{i,k} \right\|^{2} - 1 \right|^{2} + \sum_{j=i+1}^{3} \left| e_{i,k} \cdot e_{j,k} \right|^{2} \right), \tag{6}$$

where, the subscript i = 1,2,3 and k = x, y, z represent the MBC level index, and Cartesian components index, respectively. The MBC scores $\mathbf{w}_i(p)$ is regularized with a zero-mean score loss to remove the non-zero baseline:

$$L_{zms} = \frac{1}{9} \sum_{k=x,y,z} \sum_{i=1}^{3} \left| \frac{1}{N_p} \sum_{p} w_{i,k}(p) \right|, \tag{7}$$

where N_p denotes the total number of projections.

In addition, A DVF self-consistency regularization loss is included. This regularization ensures that applying the inverse DVFs $d^{-1}(x,p)$ to the deformed (dynamic) CBCTs I(x,p) brings them back to the reference CBCT $I_{ref}(x)$, thus helping address the ill-posed nature of the reconstruction problem:

$$L_{SC} = \frac{1}{N_{batch}} \sum_{N_{batch}} \frac{1}{N_{voxel}} \sum_{N_{voxel}} \left| \boldsymbol{I}'(\boldsymbol{x}, p) - \boldsymbol{I}_{ref}(\boldsymbol{x}) \right|^2, \tag{8}$$

where $I'(x,p) = I(x + d^{-1}(x,p),p)$. The inverse DVFs were calculated using a fixed-point iterative algorithm with three iterations ²⁷. In this stage, the projection similarity loss $L_{sim}^{prj}(I(x,p))$ is defined on the dynamic CBCTs I(x,p):

$$L_{sim}^{prj}(I(\boldsymbol{x},p)) = \frac{1}{N_{batch} N_{pixel}} \sum_{N_{batch} N_{pixel}} |\mathcal{P}[I(\boldsymbol{x},p)] - \boldsymbol{p}|^{2}, \tag{9}$$

In training stage II, the reference CBCT I_{ref} is first 2x upsampled to accommodate the high-resolution training. DREME follows the strategies and loss functions used in stage I to form substages II-a, II-b, and II-c. Additionally, to improve the robustness of the CNN motion encoder to varying projection angles, angular augmentation is added to substages II-b and II-c, by synthesizing DRRs at random angles from the solved dynamics CBCTs. The MBC scores inferred by the CNN motion encoder from these DRRs were compared with the known scores to calculate the angular augmentation loss. Last, a substage II-d is added to enable simultaneous angular and deformation augmentation training. Based on augmented MBC scores $w'_{i,k}(p)$, corresponding DVFs (Eq. 2) and deformation-augmented CBCTs (Eq. 1) are obtained. Randomangled DRRs are generated from these deformation-augmented CBCTs and fed into the motion model to infer the MBC scores $w''_{i,k}(p)$. The augmentation loss is defined as the sum of squared errors between the known scores $w''_{i,k}(p)$ and CNN-predicted scores $w''_{i,k}(p)$ to train a motion- and projection angle-robust CNN for real-time motion prediction:

$$L_{AUG} = \frac{1}{9N_{batch}} \sum_{N_{batch}} \sum_{k=x,y,z} \sum_{i=1}^{3} \left| w_{i,k}^{"}(p) - w_{i,k}^{"}(p) \right|^{2}. \tag{10}$$

For the original DREME framework, the training epochs for each training stage (from I-a to II-d) are 400, 700, 1700, 1000, 1000, 1000, and 1000. Detailed learning rates and loss function weightings can be found in Table 1 of the corresponding publication²³. The original DREME framework learns the reference CBCT and the motion model (including both the CNN motion encoder and B-spline-based MBCs) from scratch each time with a new set of CBCT projections, which is time consuming. However, different treatment fractions of the same patient usually share very similar anatomy and motion models. Thus, faster reconstructions of new treatment fractions could be achieved by initializing and warming up with prior reconstruction results in an adaptive manner. Based on this, we developed DREME-adapt (with two variants) and compared them with a 'cold-start' version of DREME reconstruction (DREME-cs), with details laid out in the following.

2.3 DREME-adapt workflow

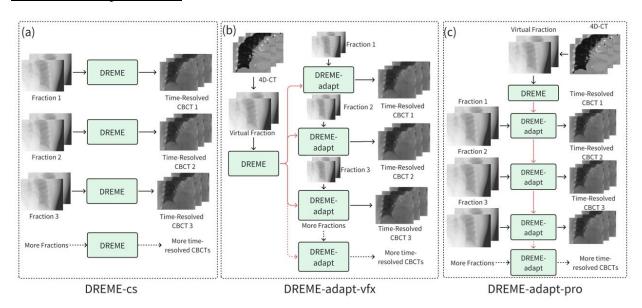


Figure 2. The workflow of three different training strategies. The virtual fraction is generated from pretreatment 4D-CT scans. The red arrows indicate DREME-adapt reconstructions initialized with the spatial INR (reference CBCT I_{ref}) and the motion model (B-spline interpolant, $e_i(x)$; and CNN-based motion encoder, $w_i(p)$) derived from previous fractions. In (a) DREME-cs, the DREME model of each fraction is trained separately from scratch. (b) DREME-adapt-vfx uses the reconstruction outputs of the 4D-CT-simulated virtual fraction as a fixed set to initialize all the subsequent reconstructions. (c) DREME-adapt-pro is a progressive, inter-fractional strategy. For each fraction, the DREME model was progressively fine-tuned from the DREME model learned from the preceding fraction (for fraction 1, it uses the DREME model derived from the 4D-CT-simulated virtual fraction) with new scan projections, forming a daisy chain.

Figure 2 illustrates the workflow of three different fractional training strategies. DREME-cs follows the original DREME framework in a 'cold-start' manner, with each fractional model trained individually and from scratch using the daily CBCT scan data. The two DREME-adapt variants, on the other hand, try to use the patient anatomy and motion correlations between the simulation and the treatment, or between different treatment fractions to greatly accelerate the training speed. For DREME-adapt-vfx, a virtual fraction of CBCT imaging is introduced, which can be simulated from a pre-treatment 4D-CT scan. The virtual fraction is trained using the original DREME model and training setup. The resultant DREME model

of the virtual fraction is then used to initialize the dynamic CBCT reconstructions of all sequential treatment fractions. Considering the potential gradual change of patient anatomy and motion patterns during the treatment, the DREME-adapt model derived from the virtual fraction may become less and less relevant in later treatment fractions. To maximize the relevance and utility of the pre-derived DREME models, the DREME-adapt-pro model, the other DREME-adapt variant, initializes each new fractional reconstruction using the model derived from the immediately preceding fraction (for fraction 1, it uses the model derived from the virtual fraction), which forms a daisy chain and allows the most relevant pre-derived DREME-adapt model to initialize each new reconstruction.

2.4 Training strategy of DREME-adapt

The DREME-adapt training strategy is based on DREME (Sec. 2.2)²³, with necessary modifications tailored to the fine-tuning task. In training stage I, the learning rates are reduced to half of those used in the original DREME method for all three substages I-a, I-b and I-c, based on the fine-tuning nature of the adaptive reconstruction approach. The epochs used for I-a and I-b are the same as the original DREME framework, allowing the model to refine the spatial INR (I_{ref}) with the new anatomy. In substage I-c, which combines motion model training, DREME-adapt uses only 700 epochs (compared to 1700 epochs for DREME) due to the similarity of motion models between different fractions. In DREME-adapt, the zero mean score regularization loss L_{ZMS} (Eq. 7) is dropped. This is because different fractions may exhibit varying motion patterns and score baselines and including L_{ZMS} could hinder the motion model's adaptation to new motion patterns.

In training stage II, the learning rates are also reduced to half of those used in the original DREME method. Additionally, the training epochs are significantly reduced. The training substages II-a, II-b and II-c use 200, 100, and 50 epochs, respectively, while all of them require 1000 epochs in the original DREME framework. Similar to DREME, the angular augmentation is added to II-b and II-c for DREME-adapt. In contrast, substage II-d is removed from DREME-adapt, as the model fine-tuned with multi-fractional data already incorporates sufficient motion variations to train a robust CNN motion encoder. In summary, DREME-adapt only uses 8.75% of the epochs needed for DREME in the training stage II, leading to much more efficient training. Detailed training setting of the fine-tuning-based DREME-adapt (for both DREME-adapt-vfx and DREME-adapt-pro) can be found in Table 1. The comparison DREME-cs method also used the same training parameter setting as DREME-adapt, without using any 'warm-start' initializations.

Table 1. The multi-resolution training strategy of DREME-adapt with different loss functions. The symbol ✓ indicates that the corresponding loss function is used in the specific training stage.

Training s	tage	I-a	I-b	I-c	II-a	II-b	II-c
Number of 6	Number of epochs		700	700	200	100	50
Spatial resolution (mm ³)			4×4×4			2×2×2	
Spatial INR lear	rning rate	2×10 ⁻⁴	2×10 ⁻⁵	5×10 ⁻⁶	5×10 ⁻⁴	2×10 ⁻⁴	5×10 ⁻⁵
Motion encoder learning rate		0.	0.	5×10 ⁻⁴	0.	0.	5×10 ⁻⁵
B-spline interpolant learning rate		0.	0.	5×10 ⁻⁴	0.	0.	5×10 ⁻⁵
Loss function W	eighting factor λ						
L_{sim}^{im}	1.	√			✓		

L_{sim}^{proj}	1.	✓	✓	\checkmark	✓	
L_{TV}	2×10 ⁻⁴	✓	✓	✓	\checkmark	
L_{MBC}	1		✓		\checkmark	
L_{SC}	1×10 ³		✓		\checkmark	
L_{AUG}	1×10 ⁻⁴			✓	✓	

2.5 Evaluation datasets and schemes

We evaluated DREME-adapt using the extended cardiac torso (XCAT) digital phantom and a lung patient dataset from the SPARE challenge⁵. The XCAT phantom can simulate different motion scenarios and anatomical variations for quantitative evaluation. The patient study helps to evaluate the clinical application potential of DREME-adapt.

2.5.1 XCAT simulation study

We simulated an XCAT phantom to cover the thorax and upper abdomen region, with a volume of 200×200×100 voxels and a 2×2×2 mm³ voxel size. A tumor was inserted into the right lung for motion tracking evaluation. Seven scenarios (C1-C7) were simulated (Table 2), each with different motion and tumor size variations. In each scenario, six fractional motion curves (one for the virtual scan fraction: VF, and five for CBCT scan fractions: F1-F5) were generated. Each motion curve covers a scan time of 60 seconds. For each VF motion curve, the motion amplitude was randomly selected from a Gaussian distribution with a mean of 20 mm and a standard deviation of 2 mm, and the motion frequency was determined by another Gaussian distribution with a mean of 0.25 Hz and a standard deviation of 0.05 Hz, and both were kept constant from cycle to cycle. The 4D-CT was generated from the first cycle by the XCAT program and then duplicated across all cycles to simulate a dynamic CBCT set with an 11-fps frame rate. For F1-F5 motion curves, inter- and intra-fractional motion frequency variations were introduced. For each curve, the motions frequency gradually increased or decreased, within the range of 0.2 Hz to 0.33 Hz (12 to 20 breath cycles per minute). For F1-F5, the motion amplitudes were drawn from the same Gaussian distribution as the VF curve, creating inter-fractional amplitude variations. B-spline-based smooth baseline shifts with a random range (lower bound set to -5 mm, and upper bound to 5 mm) were added to these motion curves to mimic the patient breathing baseline shift during the scan. For each scenario, in addition to the fractional motion variations, the tumor was modeled with a shrinking or expanding diameter ranging from 22 mm to 30 mm with a step size of 2 mm, to simulate anatomical changes and to test the robustness of the model and training strategy, though tumor expansion is less likely in clinical practice. Table 2 summarizes the inter- and intra-fractional variations of motion and tumor sizes of the seven XCAT motion scenarios (C1-C7). For each fractional motion curve (F1-F5) of each motion scenario, the 'ground-truth' dynamic CBCT volumes were generated according to the curve at a frame rate of 11 fps, yielding 660 volumes per motion curve. In clinical setups, small rigid misalignment (in addition to anatomical deformations) between 4D-CT and CBCT imaging may occur. To simulate such a potential misalignment and evaluate its effect on DREME-adapt, a rigid translation of vector length d=2mm is randomly applied to the XCAT volumes along a 3D arbitrary direction between the 4D-CT-based virtual fraction and the other five CBCT fractions. From the virtual and fractional dynamic CBCT sets, we simulated X-ray projections p from them with a rotating gantry (6°/s) in full-fan mode, covering a 360° gantry angle. The X-ray projections were acquired at 11 fps to match with the dynamic CBCT frame rate, so that each dynamic

CBCT corresponded to an X-ray projection. Each projection has an image size of 256×192 pixels with a 1.55×1.55 mm² pixel resolution. The tomographic package ASTRA toolbox was utilized to simulate the cone-beam projections²⁸, which was also used in the DREME-adapt networks as the cone-beam projector.

Table 2. Summary of inter- and intra-fractional variations of motion and tumor sizes of the XCAT motion scenarios. For the motion frequencies, " \downarrow " indicates a decrease from 0.33 Hz to 0.2 Hz, while " \uparrow " represents the opposite. For the tumor size, " \downarrow " indicates a diameter decrease from 30 mm to 22 mm (2 mm per fraction for fractions F1-F5), while " \uparrow " represents the opposite. "-" indicates no change.

XCAT scenarios	C1	C2	C3	C4	C5	C6	C7
Motion frequency variation (inter-fraction and intra- fraction)	-	\downarrow	\downarrow	1	1	\downarrow	1
Motion amplitude variation (inter-fraction)	random	random	random	random	random	random	random
Baseline shift	random	random	random	random	random	random	random
Tumor size (inter-fraction)	-	\downarrow	↑	\downarrow	↑	-	-

The performance of dynamic CBCT reconstruction of different methods was evaluated based on both the image quality of the solved dynamic CBCTs I(x, p) and the accuracy of the lung tumor localization. In detail, the image quality was evaluated via the mean relative error metric (RE):

$$RE = \frac{1}{N_p} \sum_{p} \sqrt{\frac{\sum_{l=1}^{N_{voxel}} || I(x_l, p) - I^{gt}(x_l, p)||^2}{\sum_{l=1}^{N_{voxel}} || I^{gt}(x_l, p)||^2}}$$
(11)

Where $I^{gt}(x,p)$ is the 'ground-truth' CBCT. The similarity index measure (SSIM)²⁹ and peak signal-to-noise ratio (PSNR)³⁰ were also evaluated between the reconstructed and 'ground-truth' CBCTs. The tumor localization accuracy was evaluated by contour-based metrics, including the center-of-mass error (COME)³¹ and Dice similarity coefficient (DSC)³². The 'ground-truth' contours were segmented using a thresholding-based method from the 'ground-truth' XCAT CBCTs. For reconstructed dynamic CBCTs, the tumor was first segmented using the same thresholding-based method from the reference CBCT I_{ref} . The segmented tumor was then propagated to other frames using the solved dynamic DVFs and evaluated against the 'ground-truth' contours.

2.5.2 Patient study

The patient dataset includes 7 patients (P1-P7) from the SPARE challenge dataset⁵. Six scans (P1-P4, P6, P7) cover the thoracic region and one (P5) covers the abdominal region. The dataset includes both full-fan (P1-P5) and half-fan (P6, P7) acquisition protocols. For each patient, a simulation 4D-CT set is available. The clinical 4D-CT set only has 10 phases with limited temporal resolution. To generate additional phases and motion states to mimic a dynamic CBCT sequence with higher temporal resolution for the virtual fraction, we employed a motion modeling approach based on volume registration and principal component analysis (PCA)-based DVF decomposition³³. Specifically, the end-of-exhale phase of 4D-CT was chosen

as the reference scan. Non-rigid volume registration, via the Elastix toolbox³⁴, was then applied between each of the remaining 4D-CT phases and the reference scan to derive the corresponding inter-phase DVFs. The set of DVFs were decomposed as PCA basis vectors with the corresponding coefficients. We chose the first three main PCA basis vectors to represent the motion. We subsequently interpolated the coefficients of these three basis vectors to generate 24 intermediate DVFs. The interpolated DVFs were applied to the reference scan to synthesize 24 more motion states, yielding 34 4D-CT phases in total. We repeated the 4D-CT phases for multiple cycles, creating a total of 340 (full-fan) or 643 (half-fan) volumes to simulate a dynamic CBCT set. Finally, we simulated cone-beam projections from the dynamic CBCT set (one projection from each CBCT), with the gantry angle changing from -90° to 110° (full-fan) or -90° to 270° (half-fan). The projection simulation geometry was matched to the actual projection set of each case (Table 3). In addition, the raw projections generated from 4D-CT-simulated dynamic CBCTs are of different intensity distributions from true cone-beam projections. To minimize their differences, a second-order polynomial regression method was applied to match the intensities of the virtual-fraction cone-beam projections to true projections. Table 3 details the acquisition protocol, number of CBCT fractions (virtual fraction not included), and sizes of projection sets, pixel, and reconstructed CBCTs for each patient case. For each patient case, there is one fraction with both fully-sampled and sparsely-sampled projection sets, with the latter extracted from the former to simulate a faster/lower-dose scan (for the other fractions, only the sparse set is available). For the fraction with both sets available, we used the sparse-sampled projection set to train DREME-adapt, and evaluated the CNN-based motion encoder on the fully-sampled projection set (with the sparse subset removed) to test its real-time motion prediction capability. We arranged the fractional sets in order that the dual-set fraction is always the last fraction. In Table 3, the shown projection number of the test projection sets for real-time prediction is after deducting the sparse subset projection number.

Table 3. CBCT imaging and reconstruction parameters for the patient dataset.

Patient ID	Number of Fractions	Vender	Scan mode	Projection set size for training ^a	Projection set size for real-time prediction ^a	Pixel size (mm²)	SAD ^b (mm) / SDD (mm)	Reconstructe d CBCT voxels
P1	3	Elekta	Full fan	512×512×340	512×512×675	0.8×0.8	1000/1536	200×200×100
P2	4	Elekta	Full fan	512×512×340	512×512×665	0.8×0.8	1000/1536	200×200×100
P3	3	Elekta	Full fan	512×512×340	512×512×672	0.8×0.8	1000/1536	200×200×100
P4	3	Elekta	Full fan	512×512×340	512×512×676	0.8×0.8	1000/1536	200×200×100
P5	5	Elekta	Full fan	512×512×340	512×512×667	0.8×0.8	1000/1536	200×200×100
P6	2	Varian	Half fan	750×1006×(673- 679)°	750×1006×1743	0.388×0.388	1000/1500	300×300×102
P7	2	Varian	Half fan	750×1006×(673- 677)	750×1006×2245	0.388×0.388	1000/1500	300×300×102

^a width (in pixel number) × height (in pixel number) × N_p (number of projections).

Since 'ground truth' is unavailable for the patient study, we evaluated the accuracy of the motion estimation by re-projecting the reconstructed CBCTs into DRRs and comparing them with the acquired projections, using motion features tracked by the Amsterdam Shroud (AS) method. Implementation details of the AS method were demonstrated in our previous work²². Briefly, for the DRRs and acquired projections, intensity gradients were calculated along the vertical (SI) direction and then added together along the horizonal direction across a region of interest with prominent motion (for instance, the diaphragm). The AS images capture the intensity variations resulting from respiration, and motion traces can be extracted from them to compare the reconstructed motion with that in the original cone-beam projections. To assess the similarity

^b SAD: source-to-axis distance. SDD: source-to-detector distance

^c For the half-fan patients, different fractions have different projection numbers for the sparsely-sampled training set

of the extracted motion traces, we computed the localization error²² and the Pearson correlation coefficient (PCC).

3. Results

3.1 The XCAT study results

Figure 3 compares reconstructed and 'ground-truth' dynamic CBCTs at one frame of fraction 5 (F5) using the different training strategies for all 7 XCAT scenarios (C1-C7), in the coronal (rows a-d) and axial (rows e-h) views. For each view, rows (a, e), (b, f), (c, g), (d, h) show the reconstruction results from DREME-cs, DREME-adapt-vfx, DREME-adapt-pro, and 'ground-truth', respectively. Due to the lack of 'warm-start' knowledge of the anatomy and motion patterns, the results from DREME-cs suffered from residual motion artifacts and incomplete reconstructions as it used the same limited training epoch setting as DREME-adapt (Table 1), resulting in blurry tumors and sternums, and noisy tissues as indicated by blue arrows. The DREME-adapt-vfx model was 'warmed-up' with the reconstruction from the virtual fraction. While the tissue appears less noisy, the edges of some structures such as the diaphragm and the tumor are blurred, as indicated by red arrows, due to accumulating changes simulated between the virtual fraction and fraction 5 (Table 2). In contrast, DREME-adapt-pro reconstructed the diaphragm and the tumor with sharp edges, leading to minimal motion-induced artifacts, thanks to its progressive, daisy chain design that minimizes the differences between the initialization anatomy/motion model and the to-be-solved ones.

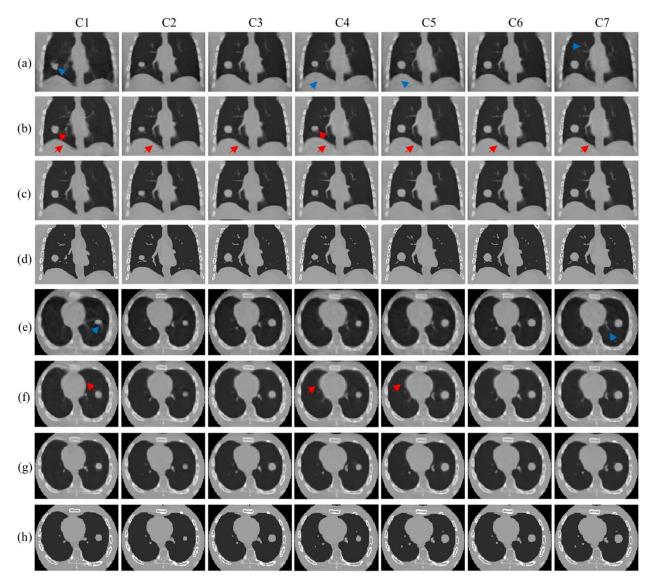


Figure 3. Comparison of reconstructed dynamic CBCTs at one frame of fraction 5, for different scenarios (C1-C7) of the XCAT study, between (a, e) DREME-cs, (b, f) DREME-adapt-vfx, (c, g) DREME-adapt-pro, and (d, h) 'ground truth'. Rows (a-d) and (e-h) show the coronal view and axial view, respectively. DREME-adapt-pro was 'warmed-up' with the anatomy/motion models solved from immediately preceding fractions for fine-tuning, leading to reconstructed dynamic CBCTs with minimal motion-induced artifacts. Blue and red arrows indicated the blurred anatomy reconstructions in DREME-cs and DREME-adapt-vfx, respectively.

To illustrate the progressive learning process of DREME-adapt-pro, figure 4 shows a frame of the dynamic CBCTs solved at different fractions (F1-F5) of scenario C2, in coronal (rows a-b) and axial (rows c-d) views. Note that the same frame may not correspond to the same motion state between F1-F5, due to daily motion variations simulated. The segmented tumors are highlighted with red contours. The edge of the diagram was well reconstructed from fraction to fraction to capture different motion states, as indicated by blue arrows. The gradual shrinkage of the tumor size was also successfully captured to match with the simulation (Table 2, C2). The progressive nature of DREME-adapt-pro's training strategy enhances its learning stability, improving its robust to anatomical and motion pattern variations.

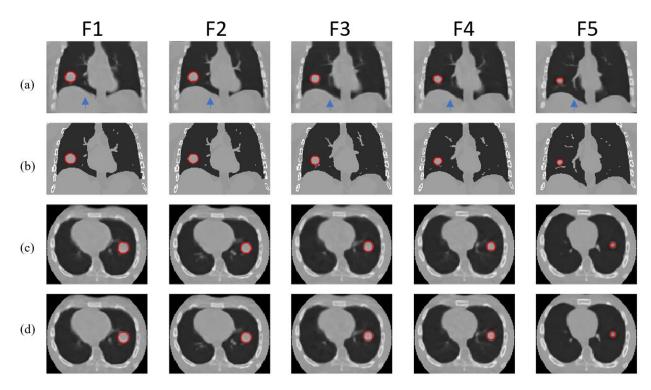


Figure 4. Comparison between (a, c) the reconstructed dynamic CBCTs and (b, d) the 'ground truth' at the same frame number, across different fractions (F1-F5) for the XCAT scenario C2. The tumor is highlighted with red contours.

The quantitative evaluation of the XCAT simulation study for all three training strategies and seven scenarios is summarized in Tables 4 and 5. The results are based on CBCT fraction 5 (F5). In addition, another strategy named DREME-nft (no fine-tuning) is also included for comparison. DREME-nft used the CNN motion encoder of the virtual-fraction DREME model to directly infer the motion and corresponding time-resolved CBCT from F5 without fine-tuning (Fig. 1), serving as a comparison baseline. Table 4 shows the metrics of dynamic CBCT reconstruction accuracy, including RE, PSNR and SSIM. DREME-adapt-pro outperformed the other strategies in terms of all three metrics. Its chain-based training helps the model to use the most relevant information for 'warming-up', resulting in more accurate anatomy details and reduced motion blur. Table 5 shows the metrics of tumor motion tracking accuracy, including COME and DSC. Overall, both DREME-adapt-vfx and DREME-adapt-pro achieved sub-voxel COME. And DREME-adapt-pro showed the best results among all strategies. DREME-nft shows better results than DREME-cs in some scenarios, highlighting the benefits of prior knowledge even under anatomy and motion changes. All Wilcoxon signed-rank tests of the five metrics comparing DREME-adapt-pro with each of the other three methods yielded p values < 10^{-3} .

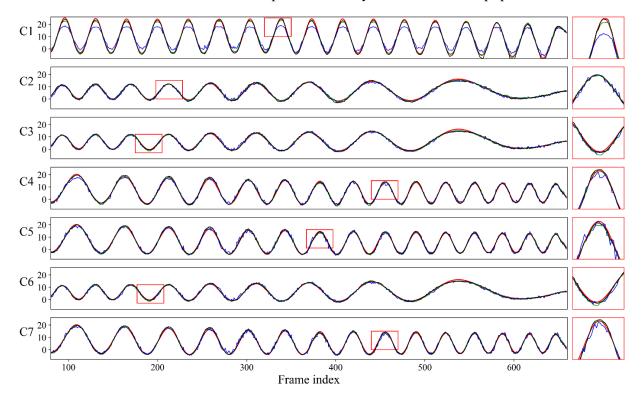
Table 4. Relative error (RE), structural similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR) of reconstructed dynamic CBCTs of the XCAT simulation study. The results for each XCAT scenario (C1-C7) are presented as the mean and standard deviation (Mean \pm SD). All Wilcoxon signed-rank tests of the three metrics comparing DREME-adapt-pro with each of the other three methods yielded p values $< 10^{-3}$.

	DREME-nft	DREME-cs	DREME-adapt-vfx	DREME-adapt-pro
XCAT	RE ↓	RE ↓	RE ↓	RE↓
scenarios	SSIM ↑	SSIM ↑	SSIM ↑	SSIM ↑
	PSNR↑	PSNR↑	PSNR↑	PSNR↑
	0.23 ± 0.10	0.21 ± 0.01	0.16 ± 0.01	0.15 ± 0.01
C1	0.941 ± 0.033	0.933 ± 0.006	0.953 ± 0.005	0.961 ± 0.006
	28.36 ± 3.80	28.42 ± 0.69	30.42 ± 0.69	31.14 ± 0.80
	0.16±0.03	0.17 ± 0.01	0.15 ± 0.01	0.14 ± 0.01
C2	0.963±0.011	0.954 ± 0.004	0.964 ± 0.002	0.969 ± 0.004
	30.85 ± 1.40	30.14 ± 0.56	31.51±0.42	32.11 ± 0.60
	0.16±0.03	0.17±0.01	0.15±0.01	0.14±0.01
C3	0.963 ± 0.011	0.953 ± 0.004	0.964 ± 0.003	0.969 ± 0.004
	31.00 ± 1.42	30.10 ± 0.56	31.44 ± 0.43	32.04 ± 0.60
	0.18±0.06	0.19±0.02	0.16±0.02	0.14±0.02
C4	0.957±0.019	0.943 ± 0.007	0.957 ± 0.007	0.967 ± 0.008
	30.09 ± 2.30	29.32±0.79	30.75 ± 0.91	31.84±1.06
	0.17±0.06	0.19±0.02	0.16±0.02	0.14±0.02
C5	0.958±0.019	0.943 ± 0.007	0.959 ± 0.007	0.966 ± 0.008
	30.23 ± 2.31	29.27±0.83	30.96 ± 0.97	31.72 ± 1.08
	0.16±0.03	0.17±0.01	0.15±0.01	0.14±0.01
C6	0.964 ± 0.011	0.953 ± 0.004	0.964 ± 0.002	0.969 ± 0.003
	30.98 ± 1.42	30.10 ± 0.56	31.42 ± 0.41	32.09 ± 0.54
	0.17±0.05	0.19±0.02	0.15±0.02	0.14±0.02
C7	0.958 ± 0.019	0.943 ± 0.007	0.959 ± 0.007	0.967 ± 0.008
	30.22 ± 2.32	29.30 ± 0.83	30.99 ± 0.99	31.77±1.07

Table 5. Center-of-mass error (COME) and Dice coefficient (DSC) as measures of solved tumor motion accuracy for the XCAT simulation study. The results for each XCAT scenario (C1-C7) are presented as the mean and standard deviation (Mean \pm SD). All Wilcoxon signed-rank tests of the two metrics comparing DREME-adapt-pro with each of the other three methods yielded p values < 10^{-3} .

XCAT	DREME-nft	DREME-cs	DREME-adapt-vfx	DREME-adapt-pro
scenarios	COME (mm) ↓	COME (mm) ↓	COME (mm) ↓	COME (mm) ↓
scenarios	DSC ↑	DSC ↑	DSC ↑	DSC ↑
C1	5.44 ± 7.06	3.98 ± 2.18	1.31 ± 0.79	1.16 ± 0.68
CI	0.737 ± 0.287	0.781 ± 0.059	0.907 ± 0.039	0.923±0.033
C2	1.19±1.63	1.06±0.58	0.82 ± 0.48	0.76 ± 0.33
C2	0.589 ± 0.030	0.904 ± 0.030	0.899 ± 0.020	0.914±0.025
C3	1.10±1.57	1.26 ± 0.94	0.87 ± 0.46	0.79±0.51
C3	0.922 ± 0.070	0.569 ± 0.011	0.939 ± 0.017	0.942 ± 0.018
C4	2.15±3.29	2.50±1.90	1.25±1.03	0.97±0.78
C4	0.569 ± 0.078	0.431 ± 0.025	0.862 ± 0.052	0.895 ± 0.039
C5	2.05±3.19	2.22±1.73	1.14±0.94	1.01±0.79
C5	0.879 ± 0.141	0.571 ± 0.023	0.924 ± 0.039	0.929 ± 0.032
00	1.13±1.56	1.24±0.94	0.87±0.46	0.72±0.45
C6	0.921 ± 0.069	0.570 ± 0.011	0.939 ± 0.017	0.947 ± 0.016
C7	2.05±3.21	1.45±1.15	1.13±0.97	1.01±0.79
C7	0.879 ± 0.142	0.910 ± 0.050	0.925 ± 0.040	0.934±0.033

Figure 5 compares the solved tumor motion trajectories of different strategies with the 'ground truth' for scenarios C1-C7 of the XCAT study, along both (top) superior-inferior (SI) and (bottom) anterior-posterior (AP) directions. The differences of trajectories are highlighted with insets for each scenario. Overall, DREME-adapt-pro showed more accurate and stable performance. DREME-cs showed a distorted curve, due to the low quality of the motion model. The DREME-adapt-vfx showed relatively smoother curves, which however contain more errors around peaks and valleys than DREME-adapt-pro.



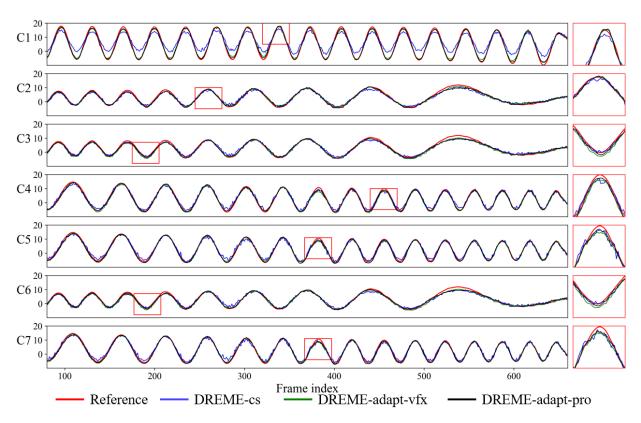


Figure 5. Comparison between solved tumor motion trajectories of three training strategies and the 'ground truth' for XCAT study scenarios C1-C7, along the (top) superior-inferior and (bottom) anterior-posterior directions. The differences between trajectories are highlighted with insets.

3.2 The patient study results

Figure 6 presents the reconstructed dynamic CBCTs at one time frame of all three training strategies from the patient study in axial and coronal views, based on the last sparsely-sampled fractional projection set of each case. We also showed the reconstruction results of the original DREME model (with full training) for reference in Fig. 6 (rows: d, h, l). DREME-adapt-pro (Fig. 6 (c, g, k)) effectively reconstructed varying patient anatomy from projections acquired under different configurations (Table 3), showing its robustness across different imaging protocols. The anatomy reconstructed by DREME-cs (Fig. 6 (a, e, i)) was blurry due to the lack of prior information and insufficient training. Although DREME-adapt-vfx (Fig. 6 (b, f, j)) was trained with a virtual-fraction model for 'warm-start', it still suffers from prominent motion artifacts and blurriness, and tissue hallucinations (for instance, patient P4) as indicated by red arrows due to outdated information within the virtual fraction model.

The localization errors and Pearson correlation coefficients are summarized in Table 6, which are similarly based on the last sparsely-sampled fractional projection set of each case. To evaluate the real-time motion inference accuracy of the fine-tuned CNN motion encoders, we used the fully-sampled projection set (Table 3) for direct motion inference and evaluation (the sparse set used during training was excluded from the fully-sampled set, as described in Sec. 2.5.2). The corresponding results are shown in Table 7 to demonstrate the prediction accuracy of different models. DREME-adapt-pro can accurately infer motion from singular X-ray projections, with the predicted motion showing a high correlation (>0.92) and a small

mean localization error (2.21 mm) when compared with that directly tracked from the original cone-beam projections using the AS method. Since the localization error was calculated in the projection domain, it would be even smaller after being projected to the 3D space, considering the magnification factor in the projection space (~1.5). Wilcoxon signed-ranked tests of COME between DREME-adapt-pro and other methods were conducted across all patients and all yielded p-values < 10⁻³, confirming the statistical significance of DREME-adapt-pro's higher motion tracking accuracy. Figure 7 compares the predicted and reference SI trajectories of various tracked anatomy for the fully-sampled projection set. The results of DREME-cs were not included due to its poor performance. The inferred motion curves by DREME-adapt-pro and DREME-adapt-vfx matched with the reference, and DREME-adapt-pro had generally better matching, especially at the extrema of the curves as the green insets indicated.

Full-Fan scan

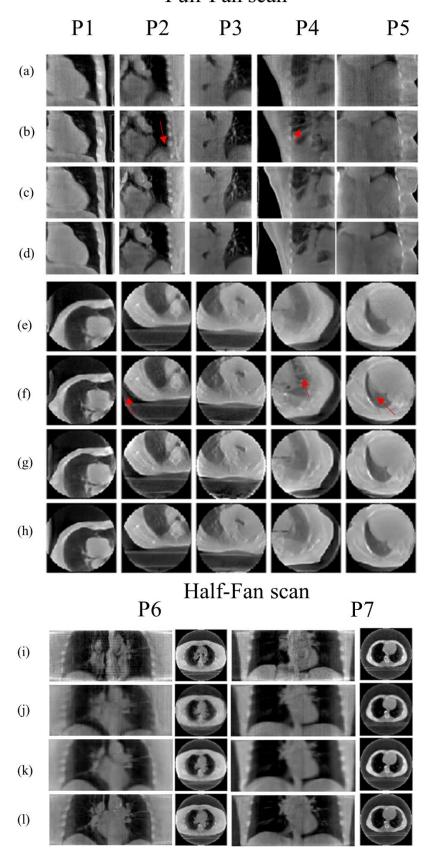


Figure 6. Dynamic CBCTs (at one time frame) reconstructed by (a, e, i) DREME-cs, (b, f, j) DREME-adapt-vfx, (c, g, k) DREME-adapt-pro, and (d, h, l) original full-training DREME for the patient study, shown in axial and coronal views. Red arrows indicated the blurred or hallucinated tissues. The model training/fine-tuning results were based on the sparsely-sampled projection set of the last fraction for each patient case.

Table 6. Pearson correlation coefficient and localization error for the SI motion trajectories extracted from the AS images, of different training strategies for the patient study. The model training/fine-tuning results were based on the sparsely-sampled projection set of the last fraction for each patient case. The localization error results were presented in terms of the mean and standard deviation (Mean \pm SD). Wilcoxon signed-ranked tests of COME between DREME-adapt-pro and other methods were conducted across all patients and all yielded p-values $< 10^{-3}$.

	DREME-cs	DREME-adapt-vfx	DREME-adapt-pro
Patient ID	Correlation ↑	Correlation ↑	Correlation ↑
	Localization error(mm) ↓	Localization error(mm) ↓	Localization error(mm) ↓
P1	0.942	0.961	0.975
PI	2.72±1.92	2.17±1.68	1.77±1.51
D2	0.932	0.966	0.975
P2	3.52 ± 3.52	2.86 ± 2.18	2.59±1.79
P3	0.849	0.923	0.932
P3	2.84 ± 2.54	2.14 ± 2.08	1.81±1.89
P4	0.936	0.959	0.961
P4	2.81±2.20	2.57 ± 1.72	2.47±1.48
D5	0.864	0.942	0.956
P5	2.79±3.09	2.37 ± 1.73	2.05±1.62
D6	0.978	0.984	0.989
P6	2.79±1.80	2.28 ± 1.91	1.89±1.56
P7	0.872	0.969	0.969
P/	4.74±4.95	2.82±1.76	2.43±2.08

Table 7. Pearson correlation coefficient and localization error for the SI motion trajectories extracted from the AS images, of different training strategies for the patient study. The results were directly obtained by testing the CNN motion encoder on the fully-sampled projection set of the last fraction for each patient case, with the CNN motion encoder being trained/fine-tuned on another sparsely-sampled projection set of the same fraction, to test its motion prediction accuracy. The results are presented in terms of the mean and standard deviation (Mean \pm SD). Wilcoxon signed-ranked tests of COME between DREME-adapt-pro and other methods were conducted across all patients and all yielded p-values $< 10^{-3}$.

	DREME-cs	DREME-adapt-vfx	DREME-adapt-pro
Patient ID	Correlation ↑	Correlation ↑	Correlation ↑
	Localization error (mm) ↓	Localization error (mm) ↓	Localization error (mm) ↓
P1	0.938	0.954	0.974
PI	2.91 ± 2.02	2.34 ± 1.79	1.81±1.52
D2	0.930	0.965	0.975
P2	3.60 ± 3.43	2.90 ± 2.18	2.61±1.79
D2	0.860	0.916	0.926
Р3	2.69±2.47	2.32 ± 2.12	2.04 ± 2.00
D4	0.930	0.962	0.961
P4	2.87±2.31	2.55 ± 1.72	2.49±1.44
D.F	0.862	0.933	0.945
P5	2.82±3.11	2.45 ± 1.94	2.11±1.89
DC	0.978	0.983	0.988
P6	2.82±1.78	2.31 ± 1.97	1.90±1.53
D7	0.866	0.968	0.969
P7	4.82±5.07	2.86±1.77	2.53±2.34

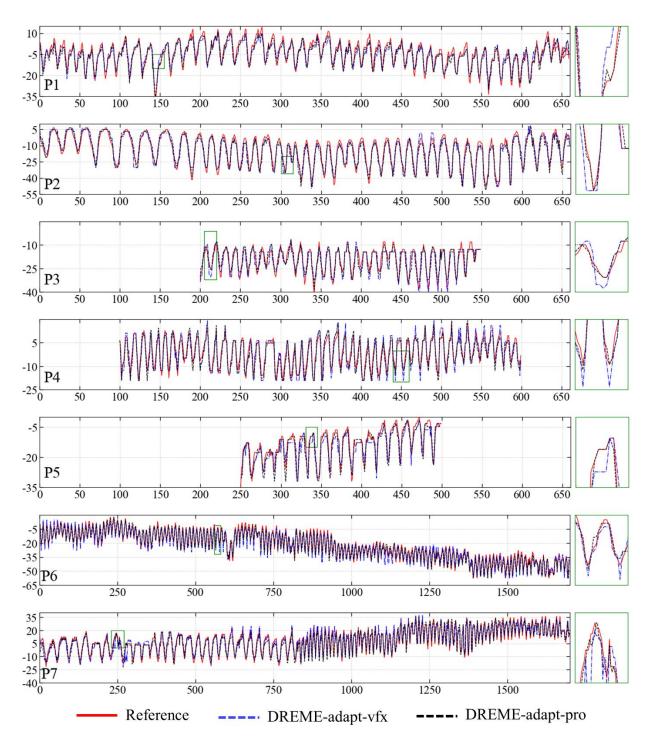


Figure 7. Comparison between the reference (red solid line), DREME-adapt-vfx (blue dashed line) and DREME-adapt-pro (black dashed line) motion trajectories extracted using the AS method for the patient study. For some patients, the tracking targets may move out of the field of view, thus only partial trajectories were available from their AS images. The results were directly obtained by testing the CNN motion encoder on the fully-sampled projection set of the last fraction for each patient case, with the CNN motion encoder being trained/fine-tuned on another sparsely-sampled projection set of the same fraction, to test its motion prediction accuracy.

4. Discussion

In this work we proposed an adaptive DREME framework for efficient time-resolved CBCT reconstruction and motion modeling in radiotherapy. The method adopted an adaptive training strategy that incorporates the DREME framework into a prior-guided reconstruction workflow. The patient anatomy and motion models solved from previous fractions are used as prior knowledge to 'warm-start' the reconstruction of following fractions. For fraction 1, due to the lack of prior treatments and CBCT scans, we used patientspecific pre-treatment 4D-CT to create a virtual dynamic CBCT sequence to generate projections to train a virtual DREME model, which serves as the prior. Based on the virtual-fraction DREME model, we evaluated two strategies: DREME-adapt-vfx, which always uses the same virtual-fraction model to 'warmstart' the reconstructions of the following fractions; and DREME-adapt-pro, which uses the virtual-fraction model to initialize the reconstruction of fraction 1, and then uses the resulting 1st fraction model to initialize the reconstruction of fraction 2, forming a progressive, daisy chain-based model. Comparing three training strategies, DREME-cs, DREME-adapt-vfx, and DREME-adapt-pro, DREME-adapt-pro showed the best performance and achieved dynamic image RE of 0.14±0.01, and tumor tracking COME of 0.92±0.62 mm for the XCAT digital phantom study. For the patient study, DREME-adapt-pro achieved a mean localization error of 2.14±1.70 mm with high Pearson correlation coefficients (>0.93) for the sparsely-sampled training dataset, and a mean localization error of 2.21±1.79 mm with similarly high Pearson correlation coefficients (>0.92) for the fully-sampled real-time motion inference testing dataset. The advantage of DREME-adaptpro, as compared with the other methods, is the use of the most relevant prior information to initialize the model training via the daisy chain, rendering it more robust to gradual anatomical and motion pattern changes through a treatment course. The training time of DREME-adapt-pro is 11 mins, only 15% of the original DREME training.

An important aspect of the DREME-adapt-pro framework is the tradeoff between the image quality and the training time. The total training could be divided into two stages: a low-resolution stage, and a highresolution stage. While the low-resolution stage is efficient in motion model learning, the reference CBCT lacks many detailed features in this stage. The high-resolution stage can refine the details of the reference CBCT image, but the reference CBCT/motion model combined training for a high spatial resolution is more time-consuming. In our full-fan patient case, the 50 epochs of stage II-c took ~2.5 mins, which is 22.7% of the total training time cost. As a trade-off that minimizes the accuracy reduction, the iterative DVF selfconsistency regulation loss (Eq. 8) could potentially be removed, which saves around 30 seconds. However, the accuracy of inverse DVFs could be crucial in applications like dose warping and accumulation³¹, in which scenarios the inverse-consistency loss function could be highly desired. Other more fundamental strategies, for instance using more efficient image representation techniques like Gaussian representation³⁵, could help to further accelerate the reconstruction speed without compromising on the reconstruction accuracy. The progressive chain-based fine-tuning strategy is also limited by the hash encoder³⁶. Because of its discontinuous nature (as compared to Fourier encoding²¹), the fine-tuning of hash encoder is not as smooth as multi-layer perceptron or convolutional layers, leading to salt-and-pepper noise in the CBCT images if insufficient epochs are used for fine-tuning. A solution to this problem is to use another differentiable encoder such as rotary position embedding^{37,38}, while the corresponding encoding efficiency gain/loss must be evaluated.

Due to the chain structure of DREME-adapt-pro, avoiding model overfitting is important; otherwise, it may reduce the reconstruction accuracy/efficiency of the next fraction. In the current framework, we used a small number (3) of MBCs $e_i(x)$ to ensure the low-rank, smooth representation of respiratory motion²⁴.

We used a 24×24×24 grid of control points for the highest B-spline interpolant to keep a balance between satisfactory motion modeling accuracy and a low risk of overfitting. However, this approach poses challenges in accurately characterizing other motion types, such as the heart beating or random, discontinuous motion caused by coughing or swallowing. To fit the more complex cardiac motion, more MBC levels may be needed³⁹. Additionally, the B-spline interpolation could be a computational bottleneck due to the large neighborhood size and substantial memory access requirements. Alternatively, adaptive sampling-based interpolation⁴⁰ or convolution-based method⁴¹ could be investigated in future for efficient interpolation.

Additionally, the training time of the DREME-adapt-pro could be further reduced by utilizing more efficient fine-tuning techniques. We re-train all parameters of the network in this study. Typical parameter-efficient fine-tuning strategies automatically optimize part of the parameters⁴² or modify the network structure by attaching additional layer/parameters while freezing the original parameters^{43,44} to use significantly less GPU memory and achieve faster training, which warrants future investigations to further accelerate DREME-adapt-pro for clinical adoption.

5. Conclusion

We proposed a prior-adapted progressive time-resolved CBCT reconstruction framework (DREME-adapt-pro) based on DREME. The proposed method first reconstructs 'virtual' dynamic CBCTs from a pre-treatment 4D-CT and then uses the derived reference anatomy and motion model for sequential fractions in a progress, adaptive manner. The results demonstrated that the framework yields high-quality motion models with sub-voxel accuracy and a ~85% reduction in training time in both simulation and real patient studies. The method could have broad applications in radiotherapy for daily anatomy tracking and motion-based plan adaptation.

Acknowledgements

The study was supported by funding from the National Institutes of Health (R01 CA240808, R01 CA258987, R01 CA280135, R01 EB034691), and from Varian Medical Systems. We would like to thank Dr. Paul Segars at Duke University for providing the XCAT phantom for our study.

References

- 1. Alderliesten T, Sonke J-J, Betgen A, van Vliet-Vroegindeweij C, Remeijer P. 3D surface imaging for monitoring intrafraction motion in frameless stereotactic body radiotherapy of lung cancer. *Radiotherapy and Oncology*. 2012;105(2):155-160.
- 2. Liu H, Schaal D, Curry H, et al. Review of cone beam computed tomography based online adaptive radiotherapy: current trend and future direction. *Radiation Oncology*. 2023;18(1):144.
- 3. Shirato H, Seppenwoolde Y, Kitamura K, Onimura R, Shimizu S. Intrafractional tumor motion: lung and liver. *Seminars in Radiation Oncology*. 2004;14(1):10-18. doi:10.1053/j.semradonc.2003.10.008

- 4. Rit S, Nijkamp J, van Herk M, Sonke JJ. Comparative study of respiratory motion correction techniques in cone-beam computed tomography. *Radiotherapy and Oncology*. Sep 2011;100(3):356-359. doi:10.1016/j.radonc.2011.08.018
- 5. Shieh CC, Gonzalez Y, Li B, et al. SPARE: Sparse view reconstruction challenge for 4D cone beam CT from a 1 min scan. *Medical physics*. 2019;46(9):3799-3811.
- 6. Chen GH, Tang J, Leng S. Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Med Phys.* Feb 2008;35(2):660-3. doi:10.1118/1.2836423
- 7. Li T, Koong A, Xing L. Enhanced 4D cone-beam CT with inter-phase motion model. *Medical Physics*. Sep 2007;34(9):3688-3695. doi:10.1118/1.2767144
- 8. Zhang Y, Yin FF, Segars WP, Ren L. A technique for estimating 4D CBCT using prior knowledge and limited angle projections. *Medical physics*. 2013;40(12):121701.
- 9. Zhang Y, Ma J, Iyengar P, Zhong Y, Wang J. A new CT reconstruction technique using adaptive deformation recovery and intensity correction (ADRIC). *Medical physics*. 2017;44(6):2223-2241.
- 10. Yasue K, Fuse H, Oyama S, et al. Quantitative analysis of the intra-beam respiratory motion with baseline drift for respiratory-gating lung stereotactic body radiation therapy. *Journal of Radiation Research*. 2022;63(1):137-147. doi:10.1093/jrr/rrab098
- 11. Tong F, Nakao M, Wu S, Nakamura M, Matsuda T. X-ray2Shape: reconstruction of 3D liver shape from a single 2D projection image. IEEE; 2020:1608-1611.
- 12. Ying X, Guo H, Ma K, Wu J, Weng Z, Zheng Y. X2CT-GAN: reconstructing CT from biplanar X-rays with generative adversarial networks. 2019:10619-10628.
- 13. Shen L, Pauly J, Xing L. NeRP: Implicit Neural Representation Learning With Prior Embedding for Sparsely Sampled Image Reconstruction. *IEEE Transactions on Neural Networks and Learning Systems*. 2024;35(1):770-782. doi:10.1109/tnnls.2022.3177134
- 14. Zhang Z, Liu J, Yang D, Kamilov US, Hugo GD. Deep learning based motion compensation for four dimensional cone beam computed tomography (4D CBCT) reconstruction. *Medical Physics*. 2022;50(2):808-820. doi:10.1002/mp.16103
- 15. Xia W, Lu Z, Huang Y, et al. MAGIC: Manifold and Graph Integrative Convolutional Network for Low-Dose CT Reconstruction. *IEEE Transactions on Medical Imaging*. 2021;40(12):3459-3472. doi:10.1109/tmi.2021.3088344
- 16. Jiang Z, Chen Y, Zhang Y, Ge Y, Yin F-F, Ren L. Augmentation of CBCT Reconstructed From Under-Sampled Projections Using Deep Learning. *IEEE Transactions on Medical Imaging*. 2019;38(11):2705-2715. doi:10.1109/tmi.2019.2912791
- 17. Zhang Y, Lv T, Ge R, et al. CD-Net: Comprehensive Domain Network With Spectral Complementary for DECT Sparse-View Reconstruction. *IEEE Transactions on Computational Imaging*. 2021;7:436-447. doi:10.1109/tci.2021.3070184
- 18. Jin KH, McCann MT, Froustey E, Unser M. Deep Convolutional Neural Network for Inverse Problems in Imaging. *IEEE Transactions on Image Processing*. 2017;26(9):4509-4522. doi:10.1109/tip.2017.2713099
- 19. Shen L, Zhao W, Capaldi D, Pauly J, Xing L. A geometry-informed deep learning framework for ultra-sparse 3D tomographic image reconstruction. *Comput Biol Med.* Sep 2022;148:105710. doi:10.1016/j.compbiomed.2022.105710
- 20. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. NeRF: representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*. 2021;65(1):99-106. doi:10.1145/3503250

- 21. Zhang Y, Shao HC, Pan T, Mengke T. Dynamic cone-beam CT reconstruction using spatial and temporal implicit neural representation learning (STINR). *Phys Med Biol*. Feb 6 2023;68(4)doi:10.1088/1361-6560/acb30d
- 22. Shao HC, Mengke T, Pan T, Zhang Y. Dynamic CBCT imaging using prior model-free spatiotemporal implicit neural representation (PMF-STINR). *Phys Med Biol*. May 23 2024;69(11)doi:10.1088/1361-6560/ad46dc
- 23. Shao H-C, Mengke T, Pan T, Zhang Y. Real-time CBCT imaging and motion tracking via a single arbitrarily-angled x-ray projection by a joint dynamic reconstruction and motion estimation (DREME) framework. *Physics in Medicine & Biology*. 2025;70(2):025026.
- 24. Li R, Lewis JH, Jia X, et al. On a PCA-based lung motion model. *Physics in Medicine and Biology*. 2011;56(18):6009-6030. doi:10.1088/0031-9155/56/18/015
- 25. Feldkamp LA, Davis LC, Kress JW. Practical cone-beam algorithm. *Journal of the Optical Society of America A*. 1984;1(6)doi:10.1364/josaa.1.000612
- 26. Zhang Y, Ma J, Iyengar P, Zhong Y, Wang J. A new CT reconstruction technique using adaptive deformation recovery and intensity correction (ADRIC). *Med Phys.* Jun 2017;44(6):2223-2241. doi:10.1002/mp.12259
- 27. Chen M, Lu W, Chen Q, Ruchala KJ, Olivera GH. A simple fixed point approach to invert a deformation fielda). *Medical Physics*. 2007;35(1):81-88. doi:10.1118/1.2816107
- 28. Van Aarle W, Palenstijn WJ, Cant J, et al. Fast and flexible X-ray tomography using the ASTRA toolbox. *Optics express*. 2016;24(22):25129-25147.
- 29. Zhou W, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*. 2004;13(4):600-612. doi:10.1109/tip.2003.819861
- 30. Huynh-Thu Q, Ghanbari M. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*. 2008;44(13):800-801. doi:10.1049/el:20080522
- 31. Zhang Y, Yin FF, Pan T, Vergalasova I, Ren L. Preliminary clinical evaluation of a 4D-CBCT estimation technique using prior information and limited-angle projections. *Radiother Oncol*. Apr 2015;115(1):22-9. doi:10.1016/j.radonc.2015.02.022
- 32. Dice LR. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302. doi:10.2307/1932409
- 33. Zhang Y, Yin FF, Segars WP, Ren L. A technique for estimating 4D CBCT using prior knowledge and limited angle projections. *Medical Physics*. 2013;40(12)doi:10.1118/1.4825097
- 34. Klein S, Staring M, Murphy K, Viergever MA, Pluim J. elastix: A Toolbox for Intensity-Based Medical Image Registration. *IEEE Transactions on Medical Imaging*. 2010;29(1):196-205. doi:10.1109/tmi.2009.2035616
- 35. Xie J, Shao HC, Zhang Y. Time-resolved dynamic CBCT reconstruction using prior-model-free spatiotemporal Gaussian representation (PMF-STGR). *ArXiv*. Mar 28 2025;
- 36. Müller T, Evans A, Schied C, Keller A. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*. 2022;41(4):1-15. doi:10.1145/3528223.3530127
- 37. Su J, Ahmed M, Lu Y, Pan S, Bo W, Liu Y. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*. 2024;568doi:10.1016/j.neucom.2023.127063
- 38. Heo B, Park S, Han D, Yun S. Rotary Position Embedding for Vision Transformer. *Lect Notes Comput Sc.* 2025;15068:289-305. doi:10.1007/978-3-031-72684-2 17
- 39. Shao HC, Qian X, Xu G, et al. A dynamic reconstruction and motion estimation framework for cardiorespiratory motion-resolved real-time volumetric MR imaging (DREME-MR). *ArXiv*. Mar 26 2025;
- 40. Huang H, Siewerdsen JH, Lu A, et al. Multi-stage Adaptive Sampling Autofocus (MASA) with a learned metric for deformable motion compensation in interventional cone-beam CT. presented at: Medical Imaging 2023: Physics of Medical Imaging; 2023;

- 41. Rueckert D, Sonoda LI, Hayes C, Hill DLG, Leach MO, Hawkes DJ. Nonrigid registration using freeform deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*. 1999;18(8):712-721. doi:10.1109/42.796284
- 42. Shen C, Gonzalez Y, Chen L, Jiang SB, Jia X. Intelligent Parameter Tuning in Optimization-Based Iterative CT Reconstruction via Deep Reinforcement Learning. *IEEE Trans Med Imaging*. Jun 2018;37(6):1430-1439. doi:10.1109/TMI.2018.2823679
- 43. Lester B, Al-Rfou R, Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning. presented at: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing; 2021;
- 44. Liao BH, Meng Y, Monz C. Parameter-Efficient Fine-Tuning without Introducing New Latency. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Acl 2023, Vol 1.* 2023:4242-4260.