Flow Matching for Respiratory Motion Interpolation from Dual-Phase CT

Anonymous Author(s)
Affiliation
Address
email

Abstract

Accurate respiratory motion characterization is essential for thoracic radiotherapy, yet 4D-CT remains inaccessible to many centers due to infrastructure and resource requirements, in addition to concerns of increased radiation dose exposure compared to 3D-CT. This accessibility gap creates treatment disparities across healthcare systems globally. To address this disparity, we present the first application of flow matching to reconstruct respiratory dynamics from standard dual-phase CT: routine images captured at maximum inhalation and exhalation. Our method combines MedVAE compression with a teacher-forced Diffusion Transformer that learns breathing patterns from 4D-CT data. Evaluating 587 sequences from 20 patients, reconstructions maintain anatomical consistency with the strongest gains at mid-respiratory positions where uncertainty peaks. Performance gains of +2.9% MS-SSIM and -6.3% LPIPS validate that generative motion synthesis from routine imaging is feasible. This foundation enables future extensions to volumetric reconstruction and tumor tracking.

1 Introduction

1.1 Clinical Motivation and Challenges

Precision radiation therapy for thoracic and abdominal tumors requires accurate targeting while minimizing exposure to healthy tissue. This requirement becomes particularly challenging due to respiratory motion. Multi-institution audits report four-dimensional computed tomography (4D-CT) reconstruction amplitude deviations of about 2 mm during regular breathing and up to 6 mm during irregular patterns [1]. These motion uncertainties correlate with worse local control rates of lung and liver metastases treated with SBRT [2].

4D-CT characterizes motion across respiratory phases and represents the current standard for tumor motion visualization in radiation treatment planning. However, adoptability and practice patterns vary widely across institutions. 4D-CT radiation dose exposure is often two to four times higher than conventional CT [3], and typically requires longer acquisition times, motion monitoring hardware, and trained staff [4]. Protocol choices also differ by scanner, reconstruction method, and phase binning, which leads to heterogeneous motion representations [1]. These technical and resource requirements create accessibility gaps—both across institutions within health systems and globally, where many low- and middle-income countries lack access to advanced imagings [5–7]. Our goal is to close this gap with methods that work under conventional CT acquisitions without 4D availability while the field progresses toward more standardized protocols.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1.2 Our Approach and Contributions

To address these disparities, we ask whether it is possible to recover clinically reliable mid-cycle motion from routine dual-phase CT without extra dose or workflow change. Our approach infers intermediate phases from end-inspiration and end-expiration. This enables motion-aware planning without 4D-CT and can lower imaging exposure and shorten simulation time. The benefit is greatest for pediatric patients and for people with cancer predisposition syndromes such as Li–Fraumeni [8, 9].

To our knowledge, this is the first adaptation of flow matching [10, 11] for medical image interpolation between sparse temporal observations. Predictions are anchored at the measured phases to prevent drift and respect observed anatomy. This boundary-constrained approach improves temporal consistency and strengthens training signals at mid-cycle where uncertainty is typically highest.

Our implementation combines MedVAE [12] for diagnostically-aware compression with a lightweight Diffusion Transformer [13] optimized for temporal reconstruction under first-last frame conditioning. The training process leverages ground truth phases (0%, 10%, 20%, 30%, 40%, 50%) with optical flow interpolation [14] for dense temporal supervision. At inference, the model requires only the endpoint phases (0% and 50%) to generate complete respiratory motion.

We validate our approach on 587 breathing sequences from the TCIA 4D-Lung dataset [15, 16] and demonstrate improvements over standard interpolation baselines, particularly at challenging mid-cycle phases where clinical uncertainty is highest. This work represents the foundational step in a comprehensive motion modeling framework. We begin with slice-wise temporal reconstruction to establish the feasibility of flow matching for respiratory motion. Future work will extend this approach to volumetric reconstruction that incorporates information from adjacent slices and to target-aware motion modeling that captures cranio-caudal dynamics essential for treatment planning. By establishing that accurate motion completion from universally available dual-phase CT is achievable, we provide a practical pathway toward democratizing advanced motion management across diverse clinical settings without requiring protocol changes or additional radiation exposure.

2 Related Work

Respiratory motion modeling and DIR. In radiotherapy, respiratory motion is often estimated with deformable image registration (DIR). Classical variational optical flow (Horn–Schunck, TV-L1, Farnebäck) and free-form B-spline registration remain strong baselines but need careful regularization and can degrade under large deformations or in homogeneous regions [14, 17–19]. Learning-based DIR, including LungRegNet and unsupervised variants, improves pairwise alignment on respiratory-correlated datasets [20–22]. In routine planning only end-inspiration and end-expiration may be available. Pairwise DIR does not directly produce temporally consistent mid-cycle frames from dual-phase inputs. Chaining deformations to fill frames can introduce drift. Simple frame filling with linear or cubic interpolation or optical flow is common but does not guarantee anatomically plausible trajectories [14, 17, 23, 24].

Flow matching and video generation. Flow Matching (and rectified-flow variants) learns a continuous-time velocity field or probability-flow ODE that transports one distribution to another, providing an efficient alternative to stochastic diffusion sampling [10, 11, 25]. For video, diffusion-style models learn spatiotemporal dynamics directly (often in latent space) and scale to high resolution, e.g., Video Diffusion Models, Pyramidal Flow Matching, and Diffusion Transformers (DiT) [13, 26, 27].

First-last-frame conditioning (FLF2V). Conditioning on the first and last frames has emerged as a practical interface for image-to-video generation. Large systems such as Meta MovieGen and Wan2.1-FLF2V-14B demonstrate high-fidelity trajectories between two keyframes, while diffusion-based academic in-betweening (e.g., VIDIM) explores sparse-anchor interpolation [28–30]. We adapt this endpoint-aware design to medical CT, where "keyframes" are clinically acquired end-inspiratory and end-expiratory phases.

Domain-specific latent compression. Operating in a compact latent space is crucial for high-resolution medical imagery. Latent diffusion established the efficacy of autoencoder latents for scalable generation, and medical foundation autoencoders such as MedVAE preserve diagnostically relevant detail while enabling efficient training [12, 31].

Supervising mid-trajectory dynamics. To strengthen intermediate states, we draw on the spirit of teacher forcing and related sequence-training strategies (scheduled sampling, Professor Forcing), which provide ground-truth context to stabilize one-step predictions [32–34]. Conceptually related "consistency" objectives target one/few-step state agreement, complementary to our velocity-based teacher-forced supervision [35, 36]. For evaluation we use the TCIA 4D-Lung collection, a standard resource in radiotherapy motion studies [15, 16].

3 Method

3.1 Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^{H \times W}$ denote the space of 2D CT slices. We consider dual-phase inputs $\mathbf{x}_{\text{input}} = \{I^{(0)}, I^{(50)}\}$, where $I^{(0)} \in \mathcal{X}$ corresponds to end-inspiration (0% breathing phase) and $I^{(50)} \in \mathcal{X}$ corresponds to the mid-cycle 50% phase (commonly corresponding to end-expiration) acquired in 4D-CT. Our objective is to generate the missing intermediate phases $\mathbf{I} = \{I^{(10)}, I^{(20)}, I^{(30)}, I^{(40)}\} \in \mathcal{X}^4$, providing a temporally complete motion sequence between the two observed anchors.

Slice Selection Strategy. Clinical CT acquisitions consist of hundreds of axial slices that form a complete 3D thoracic volume. Rather than processing the entire volume, which is currently computationally expensive for our configuration, we select three representative axial slices capturing respiratory motion across different lung regions. As shown in Figure 1, we first identify the range where lung tissue is visible, then select slices at relative positions: inferior (Slice 0), middle (Target Slice), and superior (Slice 60). In our experiments, these correspond to slices 12, 25, and 37 respectively within the lung-containing range. This slice-based approach allows us to focus computational resources on learning high-quality temporal dynamics while maintaining anatomical coverage from the lung base to the apex.

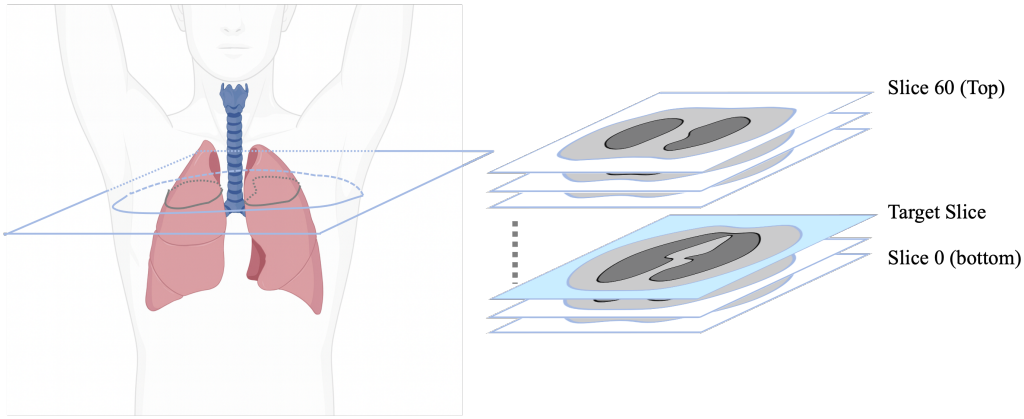

Figure 1: Anatomical slice selection strategy. From a complete thoracic CT volume, we identify slices containing visible lung tissue and select three representative levels: inferior (Slice 0), middle (Target Slice), and superior (Slice 60). This provides comprehensive coverage of respiratory motion patterns from the lung base to the apex.

3.2 Overall Architecture

Our approach combines three key components: (1) a medical VAE (MedVAE) for efficient latent space compression while preserving clinically relevant features, (2) a Diffusion Transformer (DiT) for learning respiratory motion dynamics, and (3) flow matching with teacher-forced supervision for generating smooth temporal interpolations.

3.2.1 Latent Space Compression with MedVAE

To address computational constraints while preserving medical image fidelity, we employ MedVAE with compression factor $f = 64$:

$$\mathbf{z}^{(i)} = E_\phi(I^{(i)}) \in \mathbb{R}^{H' \times W' \times C} \tag{1}$$

where E_ϕ is the MedVAE encoder with $H' = H/8, W' = W/8$. We extend the latent channels from 1 to 8 through a learned projection to provide sufficient representational capacity for motion modeling while computing KL divergence on the original single-channel distribution.

3.2.2 Respiratory Motion Modeling with DiT

We model respiratory motion using a Diffusion Transformer adapted for temporal interpolation. The architecture processes 3D latent sequences $\mathbf{z} \in \mathbb{R}^{C \times D \times H' \times W'}$ where D represents the temporal dimension. We use a configuration with hidden dimension 512 and 24 transformer layers (approximately 140M parameters).

The DiT receives conditioning from both endpoint frames [28–30]:

$$\mathbf{v}_\theta(\mathbf{z}(t), t, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}) = \mathrm{DiT}(\mathbf{z}(t), \phi_{\text{time}}(t), \phi_{\text{fff}}(\mathbf{z}^{(0)}, \mathbf{z}^{(50)})) \tag{2}$$

where $\phi_{\text{time}}(t)$ provides sinusoidal time embeddings and ϕ_{fff} encodes first-last-frame conditioning.

3.2.3 Flow Matching with Adapted Supervision

We model respiratory motion as a continuous normalizing flow between end-inspiratory and end-expiratory latent states. Let $\mathbf{z}^{(0)}$ and $\mathbf{z}^{(50)}$ denote the VAE latents of the 0% (end-inspiration) and 50% (end-expiration) phases. For curriculum and initialization we use the linear path

$$\mathbf{z}(t) = (1-t)\,\mathbf{z}^{(0)} + t\,\mathbf{z}^{(50)}, \quad t \in [0,1], \tag{3}$$

and inject a small Gaussian noise during training/inference for numerical stability. The velocity field $\mathbf{v}_\theta(\cdot, t, \mathbf{z}^{(0)}, \mathbf{z}^{(50)})$ parameterizes the ODE $\dot{\mathbf{z}}(t) = \mathbf{v}_\theta(\mathbf{z}(t), t, \mathbf{z}^{(0)}, \mathbf{z}^{(50)})$ integrated by an explicit solver.

3.2.4 Standard Flow Matching Loss:

The baseline objective minimizes:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t \sim \mathcal{U}[0,1]} \left[\left\| \mathbf{v}_\theta(\mathbf{z}(t), t, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}) - \mathbf{v}_{\text{target}}(t) \right\|_2^2 \right] \tag{4}$$

While $\mathcal{L}_{\text{FM}}$ fits the global field, it can be weak at mid-cycle times far from the two anchors (0%, 50%), where small local errors may accumulate during integration.

3.2.5 Teacher-forced supervision (local one-step consistency).

Given ground-truth latents $\{\mathbf{z}_{\text{gt}}^{(j)}\}_{j=0}^{D-1}$ at times $\{t_j\}$ with $\Delta t = t_{j+1} - t_j$, we randomly sample $j \in \{0, \ldots, D-2\}$ and enforce a one-step prediction consistency:

$$\mathcal{L}_{\text{TF}} = \frac{1}{N} \sum_j \left\| \mathbf{z}_{\text{gt}}^{(j)} + \Delta t \cdot \mathbf{v}_\theta\left(\mathbf{z}_{\text{gt}}^{(j)}, t_j, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}\right) - \mathbf{z}_{\text{gt}}^{(j+1)} \right\|_2^2 \tag{5}$$

Equivalently (up to a constant factor Δt^2):

$$\mathcal{L}_{\text{TF}} \propto \frac{1}{N} \sum_j \left\| \mathbf{v}_\theta\left(\mathbf{z}_{\text{gt}}^{(j)}, t_j, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}\right) - \frac{\mathbf{z}_{\text{gt}}^{(j+1)} - \mathbf{z}_{\text{gt}}^{(j)}}{\Delta t} \right\|_2^2 \tag{6}$$

thus providing *pointwise* constraints on the field at sparse mid-cycle times where supervision is typically scarcest. The finite difference $(\mathbf{z}_{\text{gt}}^{(j+1)} - \mathbf{z}_{\text{gt}}^{(j)})/\Delta t$ represents the true anatomical velocity between phases, capturing nonlinear tissue deformations. This direct supervision is strongest precisely where $\mathcal{L}_{\text{FM}}$'s gradient weakens (around $t \approx 0.5$), which is validated by our phase-specific results showing maximum improvement at 20-30% phases (Table 3).

3.2.6 First–Last Frame (FLF) endpoint consistency.

To keep the endpoints faithful, we add a small penalty that matches the predicted velocity to the target velocity *only at the first and last frames*:

$$\mathcal{L}_{\text{FLF}} = \tfrac{1}{2} \left\| \mathbf{v}_\theta\big(\mathbf{z}(0), 0, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}\big) - \mathbf{v}_{\text{target}}(0) \right\|_2^2 + \tfrac{1}{2} \left\| \mathbf{v}_\theta\big(\mathbf{z}(1), 1, \mathbf{z}^{(0)}, \mathbf{z}^{(50)}\big) - \mathbf{v}_{\text{target}}(1) \right\|_2^2. \quad (7)$$

Densified intermediate supervision. As detailed in Section 4.1, we augment training sequences by applying optical flow interpolation between consecutive GT phases (0%→10%, 10%→20%, etc.) to create 41-frame sequences. These densified frames provide additional supervision for $\mathcal{L}_{\text{FM}}$ and $\mathcal{L}_{\text{TF}}$, but critically differ from end-to-end optical flow (0%→50%) which would need to hallucinate all intermediate dynamics. Our approach leverages optical flow only for local temporal smoothness within 10% intervals while maintaining anatomical ground truth anchors at every respiratory phase, combining dense sampling benefits with guaranteed clinical fidelity. We emphasize that we never regress to optical flow *vectors*, and all quantitative evaluation uses exclusively ground-truth phases.

3.2.7 VAE regularization.

When the encoder/decoder are trained or fine-tuned jointly, we include a standard VAE objective to preserve anatomical fidelity in the latent space:

$$\mathcal{L}_{\text{VAE}} = \lambda_{\text{rec}}\, \mathbb{E}\Big[\|\mathbf{x} - g_\psi(f_\phi(\mathbf{x}))\|_2^2 \Big] + \beta\, \mathrm{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}))\,. \quad (8)$$

with f_ϕ the encoder and q_ϕ the approximate posterior. If the VAE is frozen during flow training, we set $\mathcal{L}_{\text{VAE}} = 0$.

3.2.8 Total objective.

The overall training objective combines the flow-matching loss, the teacher-forced local consistency loss, the first–last frame endpoint penalty, and the VAE regularization term. Each component is weighted by a scalar coefficient that controls its relative contribution:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{FM}}\, \mathcal{L}_{\text{FM}} + \lambda_{\text{TF}}\, \mathcal{L}_{\text{TF}} + \lambda_{\text{FLF}}\, \mathcal{L}_{\text{FLF}} + \lambda_{\text{VAE}}\, \mathcal{L}_{\text{VAE}}. \quad (9)$$

3.3 Training Strategy

3.3.1 Implementation Details

We employ a two-stage training strategy: (1) joint VAE-DiT training for initial learning, then (2) frozen VAE with focused DiT training. Key hyperparameters include:

- Loss weights: $\lambda_{\text{FM}} = 1.0$, $\lambda_{\text{TF}} = 1.0$, $\lambda_{\text{FLF}} = 0.1$
- AdamW optimizer with learning rate 10^{-4} and cosine annealing
- Mixed precision training (FP16) with gradient clipping at norm 1.0
- Batch size 2-4 with gradient accumulation for effective batch size 8-16
- Training on phases 0-50% (full inspiration to full expiration) for consistency
- Intensity windowing (e.g., $[-600, 900]$ HU) for lung parenchyma; rescale to $[-1, 1]$

3.4 Inference Procedure

During inference, we generate intermediate frames through ODE solving with linear initialization (rather than pure noise) for stability. The initial latent state for integration is constructed as:

$$\mathbf{z}_{\text{init}} = (1 - \alpha)\, \mathbf{z}^{(0)} + \alpha\, \mathbf{z}^{(50)} + \epsilon \quad (10)$$

where $\alpha \in [0, 1]$ specifies the target temporal position, $\mathbf{z}^{(0)}$ and $\mathbf{z}^{(50)}$ are the encoded 0% and 50% phase latents, and $\epsilon \sim \mathcal{N}(0, \sigma_{\min}^2 I)$ adds small Gaussian noise. With this parameterization, $\alpha = 0$ corresponds to initialization at the inhalation phase while $\alpha = 1$ corresponds to the 50% mid-cycle phase. We integrate the velocity field using Euler steps (50 iterations) and apply optional classifier-free guidance [37] (scale 1.5) to produce the full sequence of intermediate latents, which are then decoded by the VAE.

4 Experimental Setup

4.1 Dataset and Preprocessing

We evaluate on the 4D-Lung dataset containing 587 complete breathing cycles from 20 patients. Each cycle provides 10 respiratory phases (0%, 10%, ..., 90%) [15, 16]. We use patient-level splitting: 13 patients for training (392 cycles), 3 for validation (96 cycles), and 4 for testing (99 cycles).

Preprocessing includes: (1) Hounsfield unit windowing $[-600, 900]$ for lung tissue, (2) spatial resampling to 128×128 for computational efficiency during training (512×512 for final generation), (3) extraction of three representative axial slices per volume, and (4) phase selection focusing on 0-50% (end-inspiration to near end-expiration) for training stability.

Training Data Construction. As illustrated in Figure 2(a), we construct dense training sequences from the sparse ground truth phases. Starting with 6 GT phases from the 4D-Lung dataset (0%, 10%, 20%, 30%, 40%, 50%), we apply optical flow interpolation between consecutive phases to generate intermediate frames. This creates 41-frame sequences combining original GT frames with 35 interpolated frames. The optical flow provides smooth transitions between phases while the GT frames serve as anchor points for our teacher-forced supervision. During training, the model learns to predict velocity fields across this dense temporal sequence while being specifically constrained at GT frame positions.

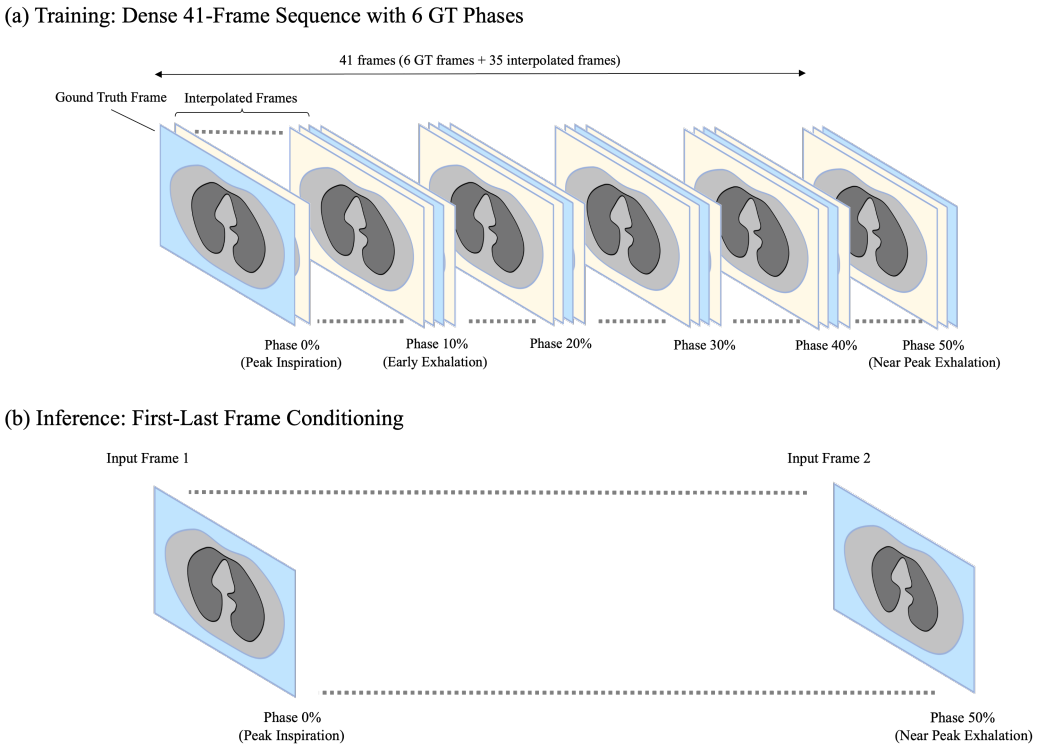

Figure 2: Training and inference data structure. (a) During training, we use dense 41-frame sequences with 6 ground truth phases from the 4D-Lung dataset, creating interpolated frames between available phases. (b) During inference, our method takes sparse dual-phase inputs (first-last frame conditioning) to generate intermediate respiratory phases.

At inference (Figure 2(b)), our method requires only the first and last frames (0% and 50%) as input. The model then generates all 39 intermediate frames through learned velocity field integration without requiring optical flow computation or access to other GT phases. For all reported metrics, supervision and evaluation use only ground-truth phases; interpolated frames are used to densify inputs but are not treated as targets.

4.2 Evaluation Metrics

We assess performance using standard image quality metrics and report mean $\pm$ std across all evaluated slices and phases:

- **MS-SSIM** (Multi-Scale Structural Similarity [38]): Measures structural preservation across multiple scales.
- **PSNR** (Peak Signal-to-Noise Ratio): Quantifies pixel-level reconstruction accuracy.
- **LPIPS** (Learned Perceptual Image Patch Similarity): Evaluates perceptual quality in deep feature space (AlexNet trunk).

We evaluate across three anatomical slice levels (inferior, middle, superior; slices 12, 25, 37) and four intermediate breathing phases (10%, 20%, 30%, 40%) to assess both spatial and temporal robustness. Error analysis includes signed error maps to identify systematic biases in intensity prediction.

4.3 Baseline Comparisons

We compare four established interpolation/registration baselines:

- **Linear interpolation:** Direct pixel-wise linear blending between frames [23, 24].
- **B-spline interpolation (Spline):** Smooth cubic spline-based frame interpolation [24].
- **Optical flow:** Dense motion field estimation using Farnebäck's algorithm [14].
- **Deformable image registration (DIR):** 2D B-spline registration from 0%→50% (SimpleITK; Mattes MI; coarse-to-fine), followed by backward warping. Intermediate phases are generated by linearly scaling the displacement field (e.g., 20% phase uses 0.4× the full deformation), which assumes linear motion between endpoints.

5 Results

5.1 Quantitative Evaluation

We evaluate four baselines (Linear, Spline, Optical Flow, DIR) across three anatomical slice levels (12, 25, 37) and four intermediate phases (10-40%), and summarize results in Table 1. Spline and Linear are comparable overall, Optical Flow trails slightly, and DIR performs substantially lower, consistent with a single endpoint deformation field failing to capture phase-dependent nonlinearity/hysteresis and with 2D registration ignoring through-plane sliding interfaces.

Table 1: Quantitative comparison on the 4D-Lung test set (mean $\pm$ std). Higher is better for MS-SSIM/PSNR; lower is better for LPIPS.

Method	MS-SSIM ↑	PSNR (dB) ↑	LPIPS ↓
Linear	0.762 ± 0.135	25.31 ± 5.10	0.103 ± 0.035
Optical Flow	0.718 ± 0.127	23.58 ± 4.57	0.111 ± 0.038
Spline	0.767 ± 0.135	25.22 ± 5.10	0.095 ± 0.035
DIR	0.660 ± 0.161	19.85 ± 2.28	0.124 ± 0.044
Flow Matching (Ours)	$\mathbf{0.789} \pm 0.138$	$\mathbf{25.80} \pm 5.12$	$\mathbf{0.089} \pm 0.035$

Compared to the strongest baseline in each metric, our method improves MS-SSIM by $+0.022$ ($\sim$2.9% vs. Spline at 0.767), increases PSNR by $+0.49$ dB (25.80 vs. Linear at 25.31 dB), and reduces LPIPS by $\sim$6.3% (0.095→0.089 vs. Spline).

5.2 Per-Slice Analysis

Performance is stable across anatomical levels, with Spline/Linear leading the classical baselines and Optical Flow slightly trailing. DIR is consistently lower across slices. Our model retains its per-slice advantage.

Table 2: Per-slice Performance (MS-SSIM / PSNR).

Method	Slice 12 (Base)	Slice 25 (Middle)	Slice 37 (Apex)
Linear	0.753 / 24.96	0.768 / 25.45	0.764 / 25.51
Optical Flow	0.709 / 23.37	0.721 / 23.67	0.724 / 23.69
Spline	0.758 / 24.87	0.773 / 25.37	0.770 / 25.43
DIR	0.647 / 19.42	0.667 / 20.04	0.666 / 20.09
Flow Matching (Ours)	**0.764 / 24.99**	**0.791 / 26.02**	**0.811 / 26.38**

5.3 Phase-Specific Analysis

All methods peak near the endpoints (10%, 40%) and degrade in the mid-cycle (20%, 30%), reflecting increased difficulty away from keyframes. DIR departs from this trend at 40%, consistent with the limitations discussed for endpoint-only deformation fields.

Table 3: Phase-specific PSNR (dB).

Method	Phase 10%	Phase 20%	Phase 30%	Phase 40%
Linear	**26.32**	24.28	24.29	26.34
Optical Flow	25.13	23.21	22.83	23.15
Spline	26.17	24.25	24.26	26.21
DIR	22.55	19.82	18.79	18.25
Flow Matching (Ours)	26.11	**25.43**	**25.26**	**26.38**

5.4 Qualitative Evaluation

Figure 3 presents visual comparisons of our flow matching interpolation across different breathing phases. We show ground truth, interpolated results, and signed error maps for phases at 10%, 20%, 30%, and 40% of the respiratory cycle.

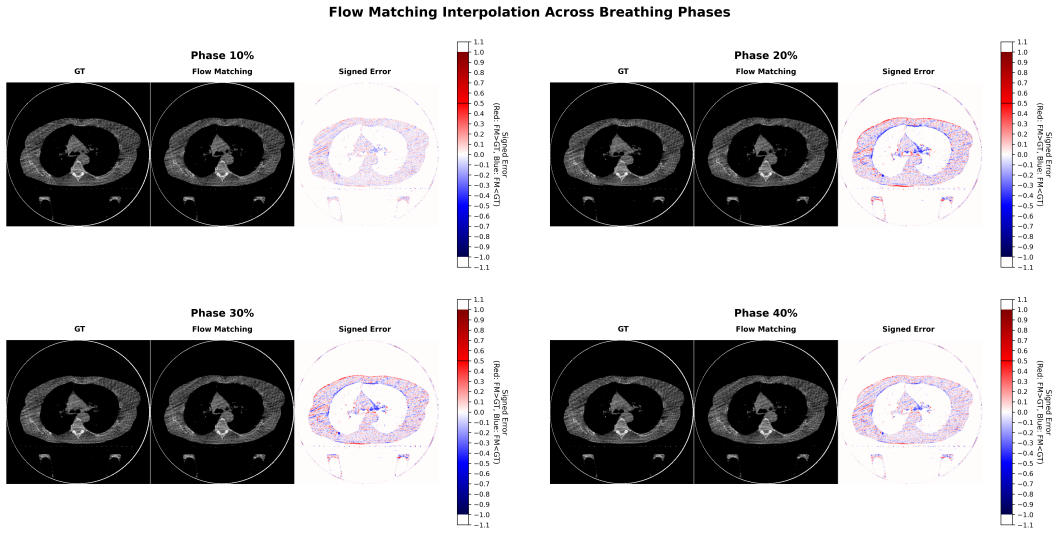

Figure 3: Flow matching interpolation across breathing phases. Ground truth (GT), flow matching results, and signed error maps (blue: under-estimation, red: over-estimation, scale: $[-1.1, 1.1]$) are shown for respiratory phases at 10%, 20%, 30%, and 40%. Anatomical structures remain well-preserved across all phases with balanced error distribution.

The signed error maps reveal balanced positive and negative deviations without systematic bias, indicating accurate intensity mapping between phases. Error magnitude follows the expected interpolation pattern for first-last frame conditioning, with minimal errors at 10% and 40% (close to

input frames at 0% and 50%) and maximum errors at 20%-30% phases where the model must infer intermediate states furthest from both conditioning frames. Notably, errors concentrate primarily at tissue boundaries and regions of respiratory motion, while homogeneous regions maintain minimal deviation. The preservation of anatomical structures across all phases, particularly at the challenging 20%-30% midpoints where uncertainty is highest, demonstrates the method's robustness and clinical viability. These visual results corroborate our quantitative findings, confirming that flow matching provides superior temporal interpolation while maintaining diagnostic image quality even at maximum interpolation distance.

5.5 Discussion

Our results demonstrate successful adaptation of flow matching to medical respiratory motion modeling. While improvements over traditional interpolation are modest ($\sim$2.9% MS-SSIM, $\sim$6% LPIPS, and +0.49 dB PSNR), they are consistent across all phases and most pronounced at challenging mid-cycle phases (20%-30%) where accurate motion characterization is clinically critical. Balanced errors and preserved structures support clinical plausibility; full clinical validation is future work.

The effectiveness stems from key domain adaptations: MedVAE preserves diagnostic features during compression, teacher-forced supervision anchors predictions to ground truth anatomy, and training on respiratory-specific phase ranges ensures physiologically realistic motion. These adaptations demonstrate that video generation techniques developed for natural images can effectively handle the unique requirements of medical imaging when properly modified.

5.6 Limitations and Future Work

This study is slice-wise and therefore lacks explicit volumetric coupling, which limits fidelity for cranio–caudal (through-plane) motion and sliding interfaces. The evaluation also focuses on regularized breathing patterns; irregular cycles and pronounced hysteresis are underexplored, and we did not assess downstream contour propagation or dose impact. Future work will extend the model to volumetric latent flows that incorporate information from neighboring slices to capture 3D consistency, and will further integrate target-aware cues (e.g., tumor or organ contours) to enable trajectory estimation and contour propagation for radiotherapy planning.

6 Conclusion

We present a novel application of flow matching to respiratory motion synthesis in medical imaging. By combining flow matching with domain-specific adaptations including medical-aware compression and teacher-forced supervision, we generate anatomically plausible intermediate respiratory phases from limited images at select phases. Our approach demonstrates that modern video generation techniques can address clinical imaging challenges when thoughtfully adapted to preserve diagnostic quality and respect physiological constraints. This work provides a foundation for applying generative modeling to medical motion, with potential impact on radiation therapy planning and delivery.

Ethics and Trust Statement

This work uses a public, de-identified dataset (TCIA 4D-Lung). The method is intended as an assistive tool; it does not replace clinical judgment. To mitigate risks from implausible motion or artifacts, we (i) constrain supervision to ground-truth phases, (ii) report signed error maps, and (iii) recommend clinician review and uncertainty visualization before downstream use. No protected health information was used.

Reproducibility Statement

We provide implementation details (architectures, losses, hyperparameters), patient-level splits, and evaluation protocols in the main text.

References

[1] J. Dhont M. Kyndt A. Gulyban J. Szkitsak E. Bogaert D. van Gestel M. Burghelea, J. Bakkali Tahiri and N. Reynaert. Results of a multicenter 4d computed tomography quality assurance audit: Evaluating image accuracy and consistency. *Physics and Imaging in Radiation Oncology*, 2023.

[2] A.K. Ozga C. Petersen L. Pinnschmidt R. Werner T. Sentker, V. Schmidt and T. Gauer. 4d ct image artifacts affect local control in sbrt of lung and liver metastases. *Radiotherapy and Oncology*, 148:229–234, 2020.

[3] T. Ishii S. Mori, S. Ko and K. Nishizawa. Effective doses in four-dimensional computed tomography for lung radiotherapy planning. *Medical Dosimetry*, 2009.

[4] J. M. Balter R. S. Emery K. M. Forster S. Jiang et al M. J. Keall, G. S. Mageras. The management of respiratory motion in radiation oncology (aapm task group 76 report). *Medical Physics*, 2006.

[5] World Health Organization. *Global atlas of medical devices*. World Health Organization, Geneva, 2022. ISBN 978-92-4-006220-7. URL https://iris.who.int/bitstream/handle/10665/364709/9789240062207-eng.pdf.

[6] Melissa Silverberg. How radiologists overcome barriers to provide imaging in low to middle income countries, July 2024. URL https://www.rsna.org/news/2024/july/imaging-in-lmics. States: "There is less than one CT scanner per million inhabitants in LMICs compared to 40 per million in high-income countries.".

[7] Emma Beede, Elizabeth Elliott Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Raumviboonsuk, and Laura Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–12, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3313831.3376718. URL https://dl.acm.org/doi/10.1145/3313831.3376718.

[8] R. K. Otto R. S. Iyer G. S. Philips J. O. Swanson C. Zacharias, A. M. Alessio and M. M. Thapa. Pediatric ct: strategies to lower radiation dose. *AJR American Journal of Roentgenology*, 2013.

[9] K. E. Nichols A. S. Levine K. Schneider, K. Zelley and J. Garber. Li–fraumeni syndrome. genereviews. *NCBI Bookshelf*, 2025.

[10] T. Karras T. Aila S. Laine Y. Lipman, R. T. Q. Chen and J. Lehtinen. Flow matching for generative modeling. In *NeurIPS*, 2022.

[11] Z. Liu T. Liu, T. Zhao and Y. Cao. Flow straight and fast: Learning to generate and refine samples in one network. In *ICLR*, 2023.

[12] et al R. Varma. Medvae: Foundation autoencoders for medical imaging, 2024.

[13] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

[14] G. Farnebäck. Two-frame motion estimation based on polynomial expansion. *SCIA*, 2003.

[15] et al K. Clark. The cancer imaging archive (tcia) – 4d lung imaging of nsclc patients, 2016. URL https://doi.org/10.7937/K9/TCIA.2016.ELN8YGLE. Dataset.

[16] et al G. Hugo. 4d lung ct in radiotherapy: Dataset and motion analysis. *Medical Physics*, 2017.

[17] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 1981.

[18] T. Pock C. Zach and H. Bischof. A duality based approach for realtime tv-l1 optical flow. *Pattern Recognition (DAGM)*, 2007.

[19] et al D. Rueckert, L. I. Sonoda. Nonrigid registration using free-form deformations: Application to breast mr. *IEEE TMI*, 18(8):712–721, 1999.

[20] et al M. Rogers. Lungregnet: Fast learning-based deformable registration for 4d lung ct. In *MICCAI*, 2020.

[21] et al A. Hering, S. Heldmann. Learn2reg: Benchmark for learning-based medical image registration. *TMI*, 2021.

[22] et al H. Xiao. A survey of deep learning-based medical image registration. *Medical Image Analysis*, 2023.

[23] R. G. Keys. Cubic convolution interpolation for digital image processing. *IEEE TAP*, 29(6): 1153–1160, 1981.

[24] T. Blu P. Thévenaz and M. Unser. Interpolation revisited. *IEEE TMI*, 19(7):739–758, 2000.

[25] et al Y. Luo, S. Zheng. Training improvements for rectified flows. In *NeurIPS*, 2024.

[26] W. Chan et al J. Ho, A. Saharia. Video diffusion models. In *NeurIPS*, 2022.

[27] et al Z. Jin, C. Yu. Pyramidal flow matching for efficient video generation. In *CVPR*, 2024.

[28] Meta AI. Moviegen: A large-scale text-to-video system, 2024.

[29] Alibaba DAMO Academy. Wan2.1-flf2v-14b-720p: First-last-frame to video generation. *Tech Report / Model Card*, 2025.

[30] et al P. Jain. Video interpolation with diffusion models. In *CVPR*, 2024.

[31] D. Lorenz P. Esser R. Rombach, A. Blattmann and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

[32] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1(2):270–280, 1989.

[33] N. Jaitly S. Bengio, O. Vinyals and N. Shazeer. Scheduled sampling for sequence prediction. In *NeurIPS*, 2015.

[34] et al A. Lamb, A. Goyal. Professor forcing: A new algorithm for training rnns. In *NeurIPS*, 2016.

[35] et al T. Song, J. Vahdat. Consistency models. In *ICML*, 2023.

[36] et al T. Song. Improved techniques for training consistency models. In *NeurIPS*, 2023.

[37] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022.

[38] E. P. Simoncelli Z. Wang and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *Proc. IEEE Asilomar Conf. Signals, Systems, and Computers*, 2003.