# Large-Scale Vision Model-Enabled Data-Efficient Image Registration for Upright Radiation Therapy

Stanford MEDICINE

Zhuoran Jiang, Yong Yang, Serdar Charyyev, Billy W Loo, Lei Xing, Xianjin Dai

*Stanford University*

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING
MEDICAL PHYSICS COMMUNITY INSPIRATION INNOVATION IMPACT

## INTRODUCTION

Upright radiotherapy offers favorable physiological and geometric conditions compared with conventional supine delivery. However, posture-induced anatomical deformation complicates the alignment of supine diagnostic images with upright simulation and treatment images, limiting the effectiveness of standard deformable image registration (DIR) methods. This work presents a data-efficient AI-enabled DIR approach designed to enable robust supine-to-upright alignment without requiring posture-matched training data.

## METHODS

The Overall workflow is shown in Fig. 1.

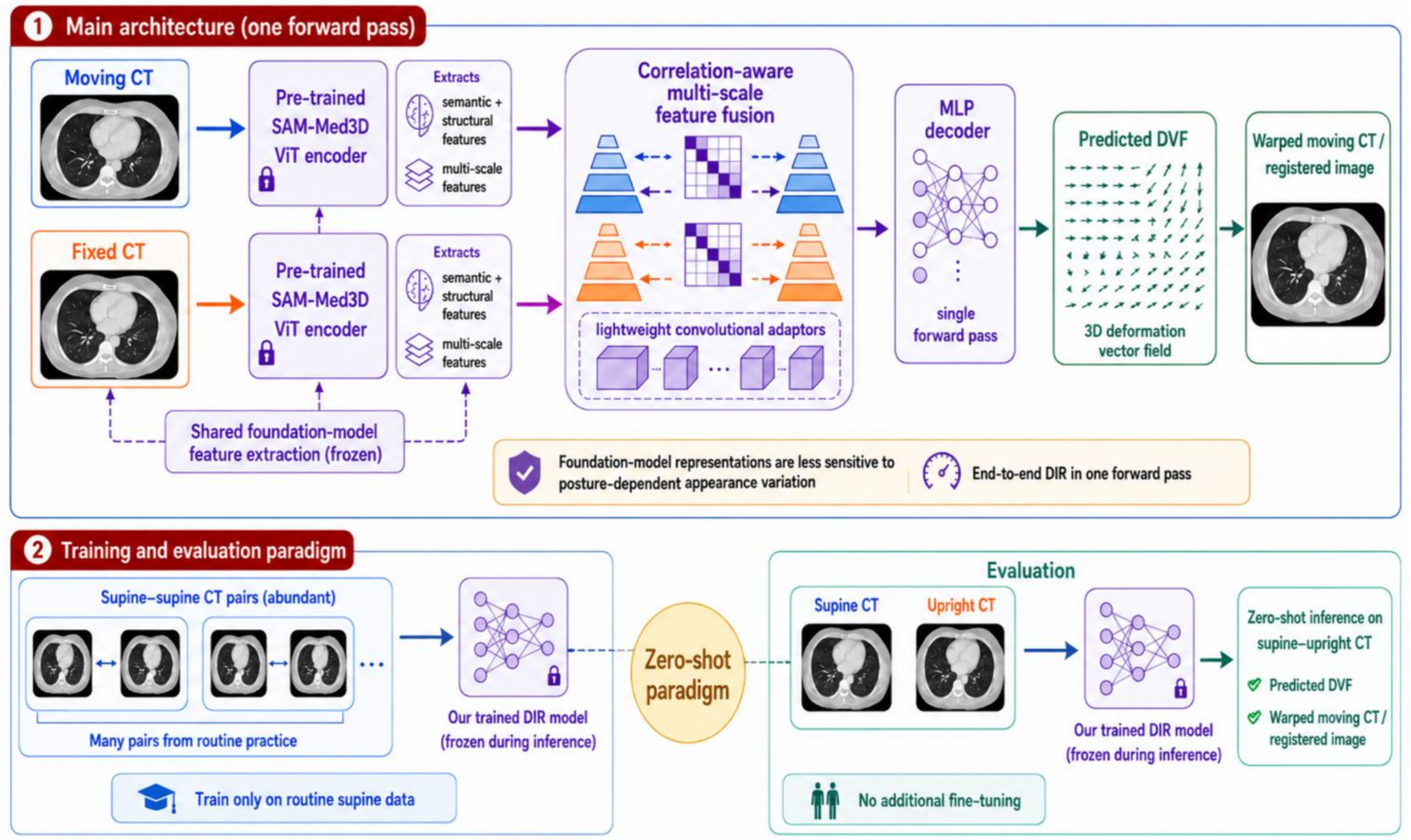


**Fig. 1. Overall workflow of the proposed large-scale vision model-enabled data-efficient image registration.**

The proposed DIR framework integrates pre-trained foundation model representations into a correlation-aware, multi-scale architecture. Leveraging large-scale vision transformer (ViT) encoders derived from SAM-Med3D, our model extracts semantic and structural features from moving and fixed images, fuses them through lightweight convolutional adaptors, and estimates deformation vector fields

## METHODS/RESULTS

(DVFs) using a multilayer perceptron (MLP) decoder. These representations are less sensitive to posture-dependent appearance variations and provide a strong basis for generalization across imaging conditions. This design enables efficient, end-to-end DVF prediction in a single forward pass, allowing the DIR framework to operate under a zero-shot paradigm. Accordingly, the model is trained exclusively using supine-supine CT image pairs, which are abundant in routine clinical practice, and subsequently evaluated on a separate supine–upright CT dataset without additional fine-tuning.

The training dataset contains 4D-CT images of lung patients scanned in a supine position. Training pairs were constructed using both intra-scan respiratory phases (inhale–exhale) and inter-scan combinations (inhale–inhale and exhale–exhale from the same patient), yielding a total of 1856 training samples. The testing dataset included lung CT scans from six patients, each with a supine diagnostic CT and a corresponding upright simulation CT. During inference, the trained model took intra-patient supine and upright CT images as inputs and predicted deformation vector fields (DVFs) to warp the supine images into the upright configuration. Registration performance was evaluated using both qualitative assessment and quantitative metrics, including root mean squared error (RMSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM). The proposed method was further benchmarked against state-of-the-art DIR models, including VoxelMorph, TransMorph, and CorrMLP.

Fig. 2 shows a representative testing case in which supine CT images were deformed and fused onto the upright CT images. All AI-based methods achieved good image similarity to the ground truth fixed images. However, due to the substantial variations in deformation patterns induced by postures, conventional deep learning models, such as VoxelMorph, TransMorph, and CorrMLP, demonstrated inferior performance in target fusion. As shown by the red and green contours in Figure 1, these models exhibited limited generalization when applied to the supine-to-upright registration task after being trained on supine-to-supine data. In contrast, the proposed foundation model-enhanced framework achieved accurate alignment of both image intensities and target structures.

This performance improvement can be attributed to the use of foundation models pre-trained on large-scale and diverse datasets, which learn semantically structured feature

## RESULTS

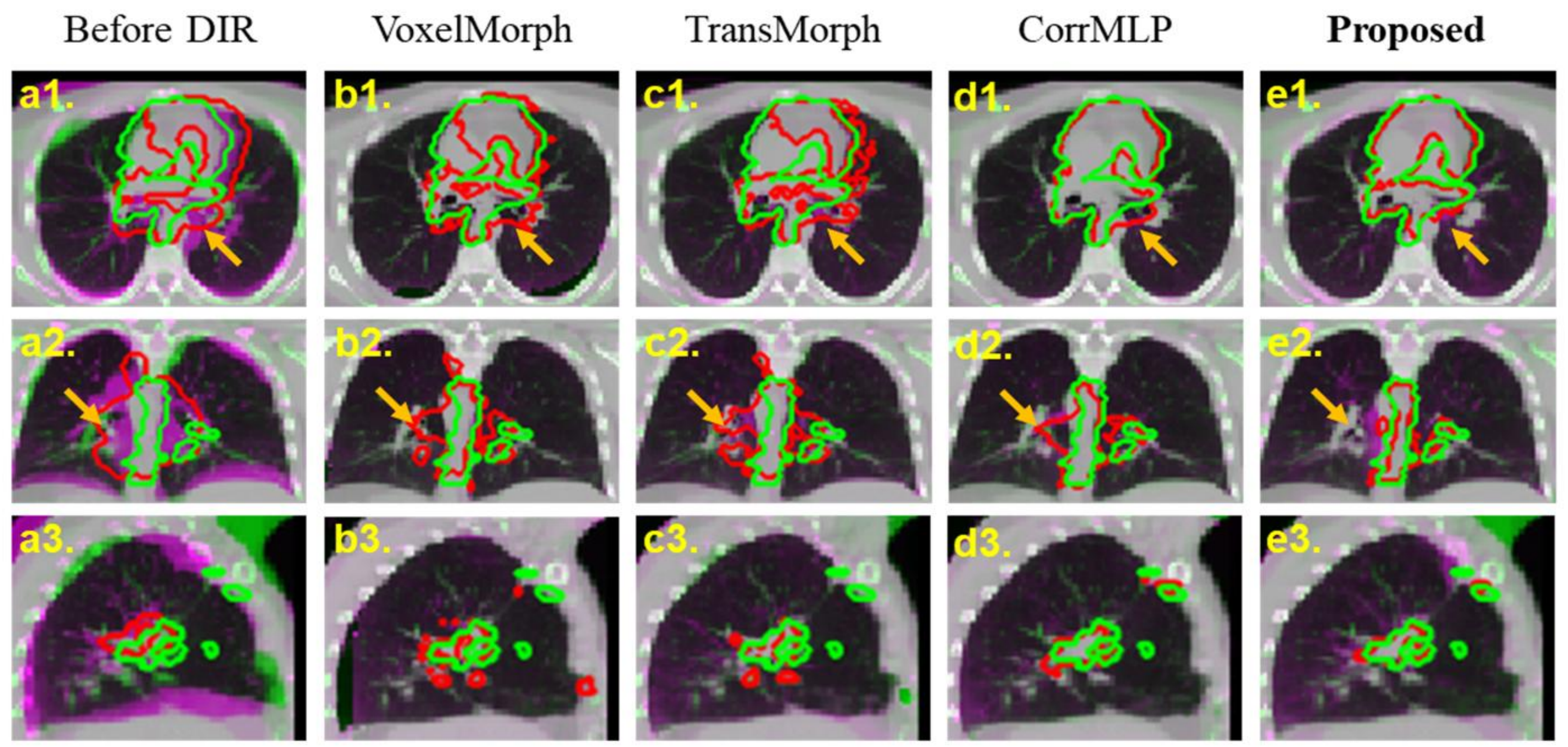


**Fig. 2. Registration result of a representative testing case. Color-overlay visualizations show the moving (red) and the fixed image (green) (a) before and (b-e) after registration by (b) VoxelMorph, (c) TransMorph, (d) CorrMLP, and (e) the proposed model, respectively. (1-3) show the axial, sagittal, and coronal views of the 3D CT data, respectively, which are displayed within range [-1000 600] HU. PTV contours are overlaid on images (red: moving; green: fixed).**

representations encoding consistent anatomical relationships beyond local image appearance. Such representations are less sensitive to variations in imaging conditions and posture-induced deformations and therefore provide a strong anatomical prior when applied within a zero-shot supine-to-upright registration framework. Quantitative analysis summarized in Table 1 further confirmed the advantages of the proposed foundation model-enhanced DIR framework.

Table 1. Quantitative analysis of registration results for all testing cases. Metrics are calculated within lung mask. Image values are normalized to [0 1].

| Metrics | Before DIR | VoxelMorph | TransMorph | CorrMLP | **Proposed** |
|---|---|---|---|---|---|
| RMSE (↓) | 0.195 | 0.068 | 0.074 | 0.055 | **0.052** |
| PSNR (↑) | 14.37 | 23.44 | 22.70 | 25.30 | **25.70** |
| SSIM (↑) | 0.264 | 0.745 | 0.702 | 0.801 | **0.806** |

## SUMMARY AND CONCLUSIONS

Large-scale foundation model enables accurate supine-to-upright DIR without requiring posture-matched supervision. These findings demonstrate the feasibility of a data-efficient, zero-shot, posture-aware DIR strategy and highlight its potential utility for image guidance and adaptive treatment planning in upright radiotherapy.