Are Video Models Emerging as Zero-Shot Learners and Reasoners in Medical Imaging?

Yuxiang Lai[1*] Jike Zhong[2*] Ming Li[3] Yuheng Li[4] Xiaofeng Yang[1,5,6†]

[1]Department of Computer Science, Emory University
[2]Department of Computer Science, University of Southern California
[3]Department of Computer Science, University of Maryland
[4]Department of Biomedical Engineering, Georgia Institute of Technology
[5]Department of Radiation Oncology and Winship Cancer Institute, Emory University

Abstract

Recent advances in large generative models have shown that simple autoregressive formulations, when scaled appropriately, can exhibit strong zero-shot generalization across domains. Motivated by this trend, we investigate whether autoregressive video modeling principles can be directly applied to medical imaging tasks, despite the model never being trained on medical data. Specifically, we evaluate a large vision model (LVM) in a zero-shot setting across four representative tasks: organ segmentation, denoising, super-resolution, and motion prediction. Remarkably, even without domain-specific fine-tuning, the LVM can delineate anatomical structures in CT scans and achieve competitive performance on segmentation, denoising, and super-resolution. Most notably, in radiotherapy motion prediction, the model forecasts future 3D CT phases directly from prior phases of a 4D CT scan, producing anatomically consistent predictions that capture patient-specific respiratory dynamics with realistic temporal coherence. We evaluate the LVM on 4D CT data from 122 patients, totaling over 1,820 3D CT volumes. Despite no prior exposure to medical data, the model achieves strong performance across all tasks and surpasses specialized DVF-based and generative baselines in motion prediction, achieving state-of-the-art spatial accuracy. These findings reveal the emergence of zero-shot capabilities in medical video modeling and highlight the potential of general-purpose video models to serve as unified learners and reasonerslaying the groundwork for future medical foundation models built on video models.

1. Introduction

Large Language Models (LLMs) and Vision-Language Models (VLMs) have fundamentally reshaped how complex problems are approached. Traditionally, medical imaging tasks required carefully engineered pipelinesfor example, using separate models for segmentation, measurement, and motion prediction. In contrast, recent medical foundation models such as MedGemma [27] demonstrate that unified frameworks can perform multiple tasks within a single system, reducing reliance on task-specific components. Similar to their general-domain counterparts, medical LLMs and VLMs are increasingly capable of tackling novel problems through few-shot in-context learning [3, 24] and even zero-shot learning, where task instructions replace fine-tuning or custom prediction heads.

Building on this trend, Wiedemer et al. [36] argue that vision models are following a similar trajectory to NLPshifting from task-specific architectures toward unified foundation models. This suggests that large-scale video models may represent the next generation of general-purpose visual backbones capable of temporal reasoning and multitask generalization.

However, general-purpose medical vision models remain largely underexplored. While task-specific architectures such as U-Net [4, 13, 26] have achieved remarkable success in segmentation, no existing framework unifies medical imaging tasks such as segmentation, registration, and motion prediction under a single paradigm. In medicine, where tasks remain highly specialized and pipelines fragmented, the potential of video models to generalize across tasks without retraining is particularly promising for building scalable and adaptive medical AI systems.

In this work, we adopt Large Vision Model (LVM) [1] as our baseline video modeling backbone. LVM is a purely visual autoregressive Transformer trained on large-scale image and video data, formulated as sequential visual sentences. It has demonstrated the ability to perform diverse visual tasks by adapting to different input patterns through prompting, without task-specific retraining, making it a

*Authors contributed equally.
†Corresponding author: xiaofeng.yang@emory.edu

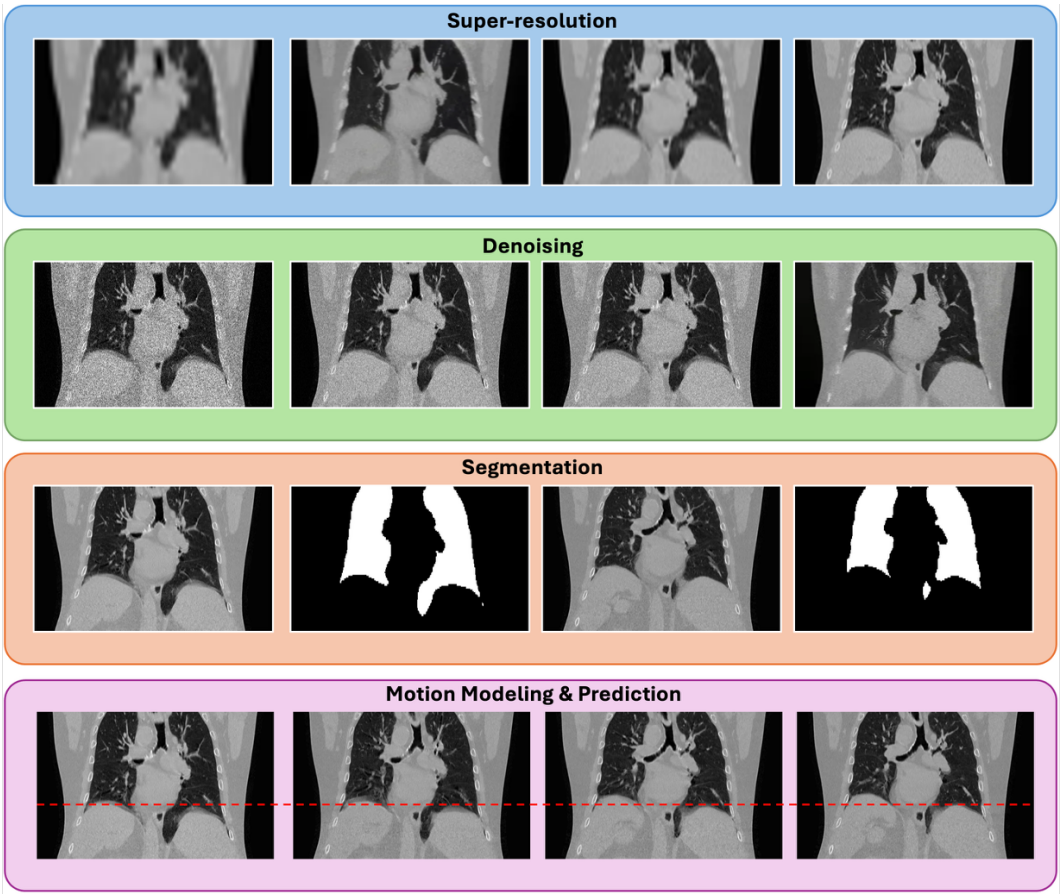

Figure 1. Zero-shot learning and reasoning examples of the video model in medical imaging. From low-level perceptual restoration (super-resolution, denoising) to high-level understanding tasks (segmentation, motion modeling, and prediction), the video model can perform a range of medical imaging tasks directly from CT sequences without task-specific training. The examples highlight the potential to further advance video models toward becoming foundational vision models for medical imaging. .

foundation for our medical video modeling experiments.

We evaluate this general-purpose video model, which has never been exposed to medical data, on five representative medical imaging tasks: segmentation, detection, denoising, super-resolution, and motion prediction. Remarkably, despite being trained entirely on natural image and video datasets, the model achieves competitive performance across these diverse tasks. In particular, for motion prediction in radiotherapy, LVM even surpasses specialized baselines by accurately forecasting future 3D CT phases based on prior ones. This improvement highlights the models ability to infer patient-specific motion patterns and reason over temporal dynamics, demonstrating strong zero-shot modeling and emergent reasoning capabilities in medical imaging. We summarize our findings as follows:

1. **Unified video models can address diverse medical imaging tasks.** Our experiments demonstrate that a single large video model can be applied to a variety of medical imaging tasksincluding segmentation, detection, denoising, super-resolution, and motion predictionwithout task-specific retraining. Although its performance does not yet reach state-of-the-art levels across all tasks, these results highlight the feasibility of developing a unified medical imaging framework through the video modeling pipeline, thereby reducing reliance on highly specialized architectures and handcrafted task designs (Figure 1).
2. **Video models are particularly effective for sequential prediction tasks.** Even in a zero-shot setting, the large vision model outperforms task-specific baselines such as deformation vector field (DVF)-based and generative models in motion prediction (Figure 2, Table 1). This finding suggests that autoregressive video models inherently capture temporal dependencies and dynamic patterns, making them well-suited for medical applications involving sequential datasuch as 4D CT, functional MRI, and dynamic ultrasoundwhere temporal coherence and motion reasoning are critical.
3. **Emergent general visual reasoning.** We observe early

signs of general visual reasoning, as the model can handle out-of-distribution (OOD) medical data and perform novel tasks without task-specific retraining. Despite never being exposed to medical images during pretraining, it generalizes across modalities and reasoning types, suggesting that large-scale video models can develop transferable representations beyond their training domains. Further investigation is needed to clarify this generalization mechanism and to assess whether pretraining on large-scale medical sequential data could further enhance reasoning and clinical applicability.

2. Related Works

Video Foundation Models. Recent progress in large-scale video modeling has shown that autoregressive and diffusion-based architectures can learn rich spatiotemporal representations directly from raw video sequences [1, 11, 37]. Inspired by the scaling trends observed in large language models (LLMs) and vision-language models (VLMs), modern video foundation models are trained with simple generative objectivessuch as next-frame prediction, masked frame reconstruction, or video continuationon massive web-scale datasets. This paradigm has led to the emergence of *zero-shot* and *few-shot* capabilities, enabling these models to perform diverse visual tasksincluding segmentation, tracking, physical reasoning, and scene manipulation-without explicit supervision [36]. Recent releases such as Veo 3 [9, 10] and Sora 2 [22] exemplify this trend, demonstrating that a single generative video model can generalize across perceptual, physical, and reasoning tasks purely through spatiotemporal sequence learning. These results suggest that large-scale video models are on a trajectory toward becoming unified, general-purpose vision foundation models, paralleling how LLMs unified natural language understanding and reasoning.

Medical Foundation Models. In the medical domain, foundation models have rapidly evolved from task-specific architectures toward unified multi-modal systems that integrate image, text, and structured data understanding [17, 20, 27, 38]. Early medical vision-language models (VLMs) such as BioViL [2] and MedCLIP [34] learned cross-modal alignment between radiology images and clinical reports, while later systems like Med-Flamingo [20] and MedGemma [27] demonstrated unified reasoning across multiple modalities and tasks, including visual question answering, report generation, and disease classification. However, despite these advances, most medical foundation models remain confined to static imaging tasks and lack explicit temporal modeling. Dynamic medical datasuch as 4D CT, MRI cine sequences, and ultrasound videoscontain complex motion patterns driven by physiological processes like respiration and cardiac cycles. Exploring whether large-scale video models can generalize zero-shot to these temporally rich modalities is thus an important next step toward true general-purpose medical AI.

Medical Motion Modeling. Organ motion modeling plays a central role in radiation therapy and interventional imaging, as it allows for precise tracking and compensation of respiratory-induced deformation. Traditional approaches rely on statistical shape and motion models, where principal component analysis (PCA) is applied to deformation vector fields (DVFs) derived from deformable image registration [21, 33]. While these models provide low-dimensional motion representations, they assume linear motion patterns and are highly sensitive to registration quality. Subsequent deep generative methods, including variational autoencoders (VAEs) [31] and diffusion models [25, 29], introduced non-linear, data-driven motion representations but still depended on precomputed DVFs for supervision, making them vulnerable to error accumulation. Sequential prediction models such as ConvLSTM [8, 28] and hybrid simulators like RMSim [18] capture temporal dynamics more effectively but remain domain-specific and require handcrafted supervision.

Zero-shot Video Models for Medical Motion. Unlike traditional methods, zero-shot video foundation models offer a unified, domain-agnostic framework for spatiotemporal reasoning. Without any retraining, these models can directly process medical 4D CT sequences as temporal input, learning to infer respiratory motion, organ deformation, and temporal coherence from natural video priors. This ability to reason over unseen modalities and capture physiological dynamics purely through autoregressive or diffusion-based mechanisms highlights the potential of video models as the next generation of medical foundation modelscapable of unifying segmentation, tracking, and motion prediction within a single zero-shot pipeline.

3. Motion Modeling and Reasoning

Overview. We formulate organ motion prediction as a video modeling and generation problem. In radiotherapy, 4D CT scans capture the dynamics of organ motion by acquiring a series of 3D CT volumes X_t at different time points t within a single respiratory cycle [12, 14]. Let $\mathcal{T} = \{0, 1, \ldots, T\}$ denote the full set of temporal indices, where T is the total number of breathing phases (typically $T = 9$ or 10). Each volume X_t corresponds to a specific breathing phase and collectively these form temporal sequence $\{X_0, X_1, \ldots, X_T\}$ that encodes organ motion.

The primary objective is to learn patient-specific motion patterns from past imaging phases and generate accurate predictions of future phases to support treatment planning and dose delivery. Rather than predicting only the next phase in isolation, the temporal trajectory itself can be mod-

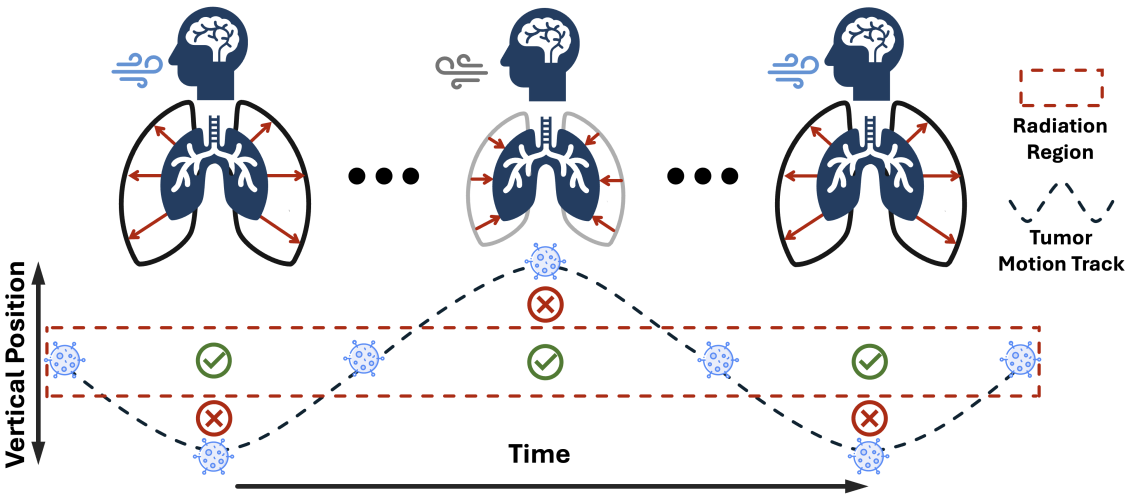

Figure 2. Schematic illustration of intrafractional tumor motion caused by respiratory cycles during thoracic and upper-abdominal radiotherapy. The top panel depicts the periodic expansion and contraction of the lungs, which drives not only pulmonary tumors but also displaces nearby organs such as the liver and heart, resulting in complex, predominantly vertical motion trajectories (blue dashed curve). The static radiation field (red dashed rectangle) is conventionally planned to encompass the entire tumor motion track to avoid geographic miss, but this inevitably irradiates more surrounding healthy tissue. The green checkmarks (✓) and red crosses (✗) mark tumor positions that are either fully covered or missed by the treatment field at different respiratory phases. This figure highlights a key clinical challenge: without accurate motion modeling, margins must be expanded to ensure target coverage for lung, liver, and cardiac-adjacent tumors, which increases unnecessary dose to adjacent organs-at-risk. Precise prediction of tumor trajectories can enable motion-adaptive strategies (e.g., gating, tracking) that safely reduce margins and support patient-specific motion management across multiple thoracic and abdominal sites.

eled as an *autoregressive process* with joint probability:

$$P(X_1, X_2, \ldots, X_T) = \prod_{t=1}^{T} p_\theta(X'_t \mid X_0, X_1, \ldots, X_{t-1}) \tag{1}$$

Here, X'_t denotes the predicted CT volume at phase t, and p_θ is the conditional probability distribution modeled by a neural network with learnable parameters θ. The network can thus make predictions conditioned on all previously observed (or generated) phases, enabling it to capture both short-term and long-range temporal dependencies in respiratory motion.

In practice, suppose we are given a partial sequence of observed phases $\{X_0, X_1, \ldots, X_L\}$, where $L < T$. We can generate the remaining future phases $\{X'_{L+1}, \ldots, X'_T\}$ by maximizing the conditional likelihood:

$$\max_\theta \prod_{t=L+1}^{T} p_\theta(X'_t \mid X_0, \ldots, X'_{t-1}) \tag{2}$$

This formulation aligns naturally with autoregressive generation tasks in natural language processing and time-series prediction, allowing the direct adoption of *autoregressive generation models* that have demonstrated strong performance in learning complex temporal dynamics. Crucially, this approach offers three core advantages: (1) It enables explicit modeling of temporal continuity and motion progression across multiple phases; (2) It allows each prediction to incorporate information from all previously observed and generated phases, promoting consistency and reducing error accumulation; (3) It directly leverages imaging data without relying on intermediate deformation vector fields, allowing the model to learn motion patterns grounded in anatomical appearances and spatial context. With this assumption, our method is illustrated as follows.

3.1. Pipeline

Pre-processing. To enable structured modeling of respiratory motion, each 3D volume in the 4D CT sequence is first segmented to isolate clinically relevant structures. We employ nnUNet [13], a state-of-the-art self-configuring segmentation framework, initialized with weights from TotalSegmentator [35]. This setup allows robust automatic segmentation of lungs, heart, and liver across varying patient anatomies and imaging conditions. nnUNet dynamically adjusts its architecture, input patch size, and training protocols based on data statistics, achieving reliable performance without manual tuning.

The resulting voxel-wise organ masks serve two essential purposes: (i) they provide geometric priors that guide motion modeling in later stages; and (ii) they serve as ground truth for evaluating prediction accuracy via segmentation-based metrics. By comparing the predicted

and reference masks across phases, we can quantitatively assess how well the model preserves organ shape, location, and boundary continuity throughout the breathing cycle.

CT tokenization. To adapt transformer-based autoregressive modeling to CT images, prior vision-language models have often divided images into fixed patches and flattened them into 1D token sequences [6]. Instead of patch embedding, LVM adopts a discrete latent representation via vector quantization (VQ) [15, 31], which encodes each CT image into a compact sequence of learnable tokens while preserving spatial structure.

As illustrated by Bai et al. [1], LVM integrates a vector-quantized generative adversarial network (VQGAN) [7] into the pipeline. VQGAN consists of a CNN-based encoder, decoder, and a discrete codebook. The encoder compresses input CT slices into a latent feature map through multiple downsampling stages; this latent grid is quantized into discrete codebook entries, yielding a 16×16 grid of tokens (256 tokens for each 256×256 CT slice). The decoder then reconstructs CT images from predicted token sequences through aligned upsampling layers. This tokenization strategy significantly reduces sequence length compared to raw pixel modeling, making autoregressive training tractable while retaining anatomical representations.

Sequence modeling of CT phases. Once tokenized, each CT phase in the 4D sequence becomes a temporally ordered grid of discrete visual tokens. LVM models this sequence using a unidirectional decoder-only transformer with causal self-attention [32]. Inspired by autoregressive language models [3], the transformer generates the token sequence for each future CT phase one step at a time, conditioned on all previously seen (or predicted) tokens.

During training, the model is exposed to the complete sequence of phases $\{X_0, ..., X_T\}$ and learns to predict the token grid of phase X_t from preceding phases $\{X_0, ..., X_{t-1}\}$. During inference, only an initial prefix ($\{X_0, ..., X_L\}$) is observed, and the model autoregressively predicts future motion phases. By feeding prior CT phases as sequential context, the transformer progressively predicts token sequences for subsequent phases, enabling multi-phase motion prediction rather than single-step prediction. This formulation allows the model to learn patient-specific respiratory patterns directly from image sequences, leading to more temporally consistent and anatomically plausible phase generation.

3.2. Implementation Details

Organ masks for the lungs, heart, and liver were extracted using the nnUNet model provided by the TotalSegmentator toolkit [35]. nnUNet is a UNet-based segmentation framework [26] that automatically configures its architecture, patch size, and training parameters based on dataset-specific characteristics [5, 16, 19]. The resulting segmentation masks were used for evaluation and as auxiliary labels to assess motion prediction accuracy.

For image tokenization, we employed a vector-quantized generative adversarial network (VQGAN) [7] with a downsampling factor of $f = 16$ and a codebook size of 8192. Each 256×256 CT slice is compressed into a 16×16 latent grid (256 tokens). These tokens are then flattened into a 1D sequence, producing a compact and discrete representation for each CT phase. Pre-trained VQGAN parameters from Yutong *et al*. [1] were adopted and fine-tuned on our 4D CT data to ensure robust encoding and reconstruction quality.

LVM uses an autoregressive modeling framework built on a decoder-only Transformer based on the LLaMA architecture [30]. To ensure reproducibility, we explicitly detail the configuration: the model comprises 12 Transformer layers, each with 16 self-attention heads and an embedding dimension of 1024. The feed-forward layers use an expansion factor of 4 (MLP hidden dimension of 4096). Rotary positional embeddings (RoPE) are applied to preserve the temporal order of tokens across phases, and causal attention masking ensures that each token is predicted only from preceding tokens. In total, the model contains approximately 120 million parameters and supports a context length of 4096 tokens, corresponding to up to 16 CT phases (16×256 tokens) within the VQGAN tokenizer framework. This architecture enables the model to capture extended intra-fraction temporal dependencies while remaining computationally tractable.

4. Experiment

4.1. Experiment Setting

Dataset. Our study utilizes two 4D CT datasets. The first dataset, provided by Hugo *et al*. [12], is publicly available and includes 4D CT scans from 20 patients diagnosed with locally advanced non-small cell lung cancer. All scans were acquired using a 16-slice helical CT scanner (Brilliance Big Bore, Philips Medical Systems, Andover, MA) with respiration-correlated CT imaging. Each scan includes 10 3D CT phases (0% to 90%) generated by phase-based binning, with a slice thickness of 3 mm.

The second dataset was collected at the Department of Radiation Oncology, Emory University. Similar to the public dataset, each 4D CT scan consists of 10 breathing phases covering a complete respiratory cycle. Scans were performed using a Siemens SOMATOM Definition AS scanner at 120 kV, following the standard Siemens Lung 4D CT protocol. The imaging system was configured with Syngo CT VA48A software, a pitch of 0.8, and a Bf37 reconstruction kernel. The image resolution is $512 \times 512 \times (133\text{-}168)$ voxels, with voxel spacing of $0.9756 \times 0.9756 \times 2.0$ mm^3 along the x, y, and z axes.

The third dataset is an in-house liver cancer therapy co-

Table 1. **Performance on Next-Phase Motion Prediction.** LVM significantly outperforms previous methods and achieves high accuracy. We evaluate performance by comparing the segmentation masks of organs between the predicted and ground-truth CT phases. Higher IoU and DSC values indicate more precise overall predictions, while lower SD and HD values suggest more accurate boundary alignment.

Data	Organ	Method	IoU (%) ↑	DSC (%) ↑	SD (mm) ↓	HD95 (mm) ↓
Public	Lung	DAM [23]	80.08	85.47	4.84	7.14
		DiffuseRT [29]	81.23	86.87	4.19	6.92
		ConvLSTM [8]	82.36	88.14	3.92	5.76
		RMSim [18]	84.05	89.16	3.66	5.08
		Ours	90.75	95.15	2.65	4.33
	Heart	DAM [23]	78.62	82.69	5.59	9.83
		DiffuseRT [29]	80.29	85.08	4.95	9.12
		ConvLSTM [8]	81.41	86.27	4.41	8.45
		RMSim [18]	82.97	87.96	3.94	7.14
		Ours	88.43	93.85	2.39	4.24
Private	Lung	DAM [23]	79.71	83.15	5.98	8.16
		DiffuseRT [29]	82.44	85.36	5.44	7.23
		ConvLSTM [8]	84.06	88.44	4.54	6.81
		RMSim [18]	85.35	89.62	3.85	6.54
		Ours	91.88	95.74	2.84	4.47
	Heart	DAM [23]	76.81	80.97	6.63	9.05
		DiffuseRT [29]	79.07	83.43	6.29	8.58
		ConvLSTM [8]	80.54	85.25	5.77	7.59
		RMSim [18]	82.14	87.03	5.32	7.15
		Ours	87.83	93.24	3.34	5.38
	Liver	DAM [23]	77.15	81.46	4.99	9.34
		DiffuseRT [29]	79.18	83.04	4.77	8.63
		ConvLSTM [8]	82.68	86.18	4.35	8.26
		RMSim [18]	83.05	88.08	3.70	6.75
		Ours	91.99	95.83	2.95	5.09

Public Lung & Heart: 80 4D CT scans, 800 3D CT scans; Private Lung & Heart: 50 4D CT scans, 500 3D CT scans; Private Liver: 52 4D CT scans, 520 3D CT scans.
IoU - intersection over union; DSC - dice similarity coefficient; SD - surface distance; HD - Hausdorff distance.

hort comprising 52 patients with hepatocellular carcinoma (HCC) who underwent pre-treatment 4D CT scans. All scans were acquired using a [scanner name, e.g., Siemens SOMATOM Force] at 120 kV, following the standard liver 4D CT protocol with respiratory phase-based binning (10 phases). The slice thickness was 2 mm, and the in-plane resolution is 512×512 with voxel spacing of [e.g., $0.976 \times 0.976 \times 2.0$ mm^3]. Patients were scanned in the supine position under free breathing. Institutional ethics approval was obtained for retrospective data use.

Evaluation. Our primary goal is to evaluate how well the zero-shot video model can perform medical imaging tasks without domain-specific training. For radiotherapy motion prediction, we focus on assessing the accuracy of predicted organ positions and shapes rather than low-level pixel similarity. Spatial agreement between predicted and ground-truth segmentations is evaluated using four widely adopted metrics: Intersection over Union (IoU), Dice Similarity Coefficient (DSC), Surface Distance (SD), and Hausdorff Distance (HD). IoU and DSC measure volumetric overlap between predicted and reference masks, with DSC being particularly informative for small or irregular structures. SD quantifies the average symmetric distance between organ surfaces, while HD represents the maximum point-to-point deviation, reflecting worst-case boundary errors. Together, these metrics provide a comprehensive view of spatial accuracy and indicate whether radiation can be delivered precisely to the target while sparing nearby healthy tissues. For each generated CT phase, organ masks are obtained using the nnUNet [13] model, ensuring consistent segmentation quality across all phases and preventing potential bias from using different tools for ground-truth and evaluation.

Table 2. **Zero-shot performance of the large video model (LVM) across representative medical imaging tasks.** Each task is evaluated using its most widely adopted quantitative metrics, reflecting the models cross-task generalization ability and its potential to inspire the development of the unified medical vision model.

Task	Metric	Zero-shot LVM
Segmentation	DSC	91.52
	IoU	87.12
Super-resolution	SSIM	78.35
	PSNR	29.56
Denoising	SSIM	85.38
	PSNR	34.44

Beyond motion prediction, we further assess the zero-shot generalization ability of the video model across three representative medical imaging tasks: *segmentation*, *denoising*, and *super-resolution*. For segmentation, the model receives CT slices and text prompts specifying target organs (e.g., segment the liver) and outputs binary masks, evaluated using Dice and IoU. For denoising, synthetically corrupted CT images are restored to clean versions, with performance measured by PSNR and SSIM. For super-resolution, low-resolution CT inputs are upscaled to high-resolution predictions, again evaluated by PSNR and SSIM to assess structural fidelity. Although the model is not trained on medical data, it achieves stable results across all three tasks, demonstrating its ability to understand anatomical structures, preserve texture continuity, and generalize across distinct medical imaging objectives.

4.2. Results & Analysis

4.2.1. Segmentation, Denoising, and Super-Resolution

We further evaluate the zero-shot capability of the large video model (LVM) on three representative medical imaging tasks: organ segmentation, image denoising, and super-resolution. These tasks collectively cover structural understanding, low-level enhancement, and spatial fidelitykey components of clinical image interpretation.

As shown in Table 2, the zero-shot LVM achieves promising performance across all tasks without any domain-specific training. For segmentation, the model produces accurate organ boundaries with a Dice score of 91.52%, demonstrating its ability to localize anatomical structures directly from CT inputs. For denoising and super-resolution, LVM attains PSNR values of 34.4 dB and 29.6 dB, respectively, with SSIM around 0.8 in both cases, indicating perceptual quality and structure preservation.

While these results remain below the best supervised or fine-tuned models, they are notable given the absence of medical data during training. The LVM demonstrates strong generalization and the ability to infer spatial priors, texture consistency, and anatomical structure in a zero-shot setting. These findings highlight the potential of video models as general-purpose medical imaging backbonescapable of transferring across tasks and modalities.

4.2.2. Next-Phase Motion Prediction.

We evaluated *LVM* against recent deep learning methods for organ motion prediction, including DAM [23], DiffuseRT [29], ConvLSTM [8], and RMSim [18]. As shown in Table 1, *LVM* consistently outperformed all baselines across both public and private datasets.

On the public dataset, *LVM* achieved an IoU of 90.75% and a DSC of 95.15% for lung motion prediction, exceeding the best baseline (RMSim) by 5.6% (IoU) and 5.1% (DSC). For heart motion, it reached an IoU of 88.43% and a DSC of 93.85%, and further improved boundary alignment, with lower SD (6.88 mm) and HD (22.04 mm) compared to RMSim (9.34 mm and 27.24 mm, respectively).

On the private dataset, which includes more challenging and heterogeneous motion patterns, *LVM* maintained strong performance. It achieved an IoU of 91.88% and a DSC of 95.74% for lung motion prediction, outperforming the best baseline by 6.3% (IoU) and 5.9% (DSC). For heart motion, *LVM* achieved an IoU of 87.83% and a DSC of 93.24%, with clear reductions in surface distance (7.38 mm vs. up to 13.25 mm for DAM) and Hausdorff distance (23.97 mm vs. 34.81 mm). For liver motiona particularly difficult organ due to its irregular deformation*LVM* achieved an IoU of 91.99% and a DSC of 95.83%, surpassing the next-best model by 8.7% (IoU) and 7.0% (DSC). These results highlight two findings: (1) *LVM* produces segmentation masks with higher spatial agreement (IoU/DSC) and more accurate boundary alignment (lower SD/HD) than previous methods; and (2) These gains are consistent across organs and datasets, demonstrating generalizability under diverse imaging conditions.

We attribute this performance advantage to the autoregressive design of *LVM*. Existing generative approaches (e.g., DAM [23], DiffuseRT [29]) often rely on limited conditioning information (typically a single CT phase), which restricts their ability to represent the full temporal dynamics of breathing. Sequence-based models like ConvLSTM [8] partially address this by propagating temporal states, while physics-based simulators like RMSim [18] embed predefined assumptions about motion fields. However, these methods either lack strong mechanisms for long-sequence consistency or are constrained by their assumptions.

In contrast, *LVM* conditions each prediction on a sequence of past CT phases and feeds its predictions into subsequent steps, creating a feedback loop. This design allows the model to catch patient-specific motion patterns and maintain temporal consistency across predicted phases,

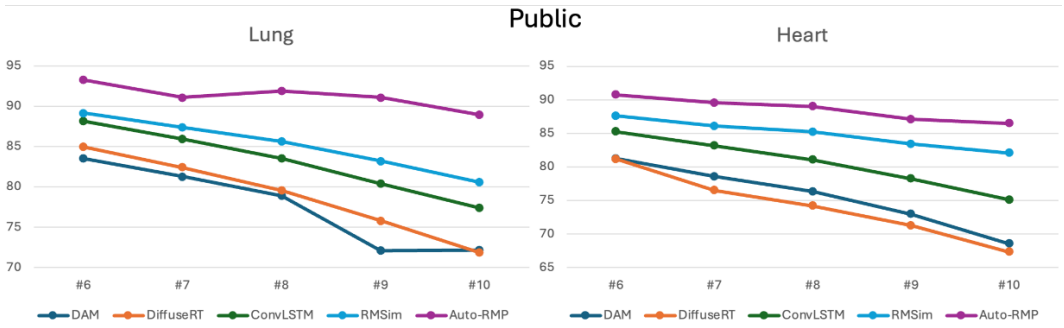

Figure 3. **Multi-phase motion prediction on the public dataset.** We evaluate model performance on the public 4D CT dataset using Dice Similarity Coefficient (DSC, %). Each model is provided with the first five phases of the 4D CT scan and autoregressively predicts the next five phases. The plots show phase-by-phase DSC for five representative methods (DAM, DiffuseRT, ConvLSTM, RMSim, and our proposed LVM). LVM consistently achieves the highest DSC across all predicted phases and exhibits the smallest performance drop from phase #1 to #5, indicating its superior ability to model smooth and realistic multi-phase motion patterns.

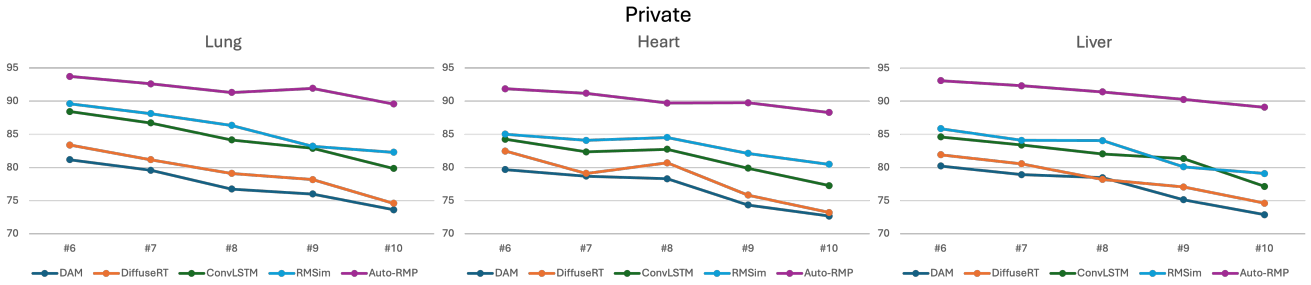

Figure 4. **Multi-phase motion prediction on the private dataset.** The same DSC-based evaluation is conducted on our institutional 4D CT dataset (including lung, heart, and liver cases). Each model receives the first five phases and must generate the subsequent five phases. LVM maintains consistently higher DSC across all organs and phases, with smoother phase-to-phase transitions and less degradation compared to competing methods, demonstrating strong robustness and generalization to in-house data.

which closely aligns with the requirements of 4D CT motion prediction.

4.2.3. Multi-phase Motion Prediction.

The multi-phase motion prediction task provides a rigorous benchmark for evaluating a video models ability to reason about temporal dynamics and maintain long-term consistency in a purely zero-shot setting. As shown in Figures 3 and 4, baseline models exhibit a steady decline in Dice Similarity Coefficient (DSC) as predictions extend across future CT phaseshighlighting the well-known challenge of error accumulation in sequential motion forecasting.

Among supervised baselines, ConvLSTM and RMSim perform better than diffusion- or GAN-based models (e.g., DiffuseRT, DAM) because they incorporate either sequential recurrence or physics-informed priors. However, these approaches depend on deformation vector field (DVF) supervision and never directly observe CT appearance, limiting their ability to learn patient-specific respiratory dynamics or spatially coherent motion patterns.

In contrast, the zero-shot LVM achieves strong temporal coherence and spatial fidelity across all datasets, maintaining DSC above 85% even at the fifth predicted phase. Despite never being trained on medical data, the model learns to infer and reason about organ motion trajectories directly from image sequencescapturing both periodic breathing patterns and smooth anatomical deformation. Its autoregressive design allows previously predicted frames to serve as temporal context, effectively reducing drift and maintaining realistic continuity over long horizons.

These findings reveal that large video models, even in zero-shot settings, can exhibit emergent abilities in medical reasoningintegrating visual understanding, memory, and prediction across multiple phases. This suggests that video models hold strong potential as unified, general-purpose backbones for medical video analysis, capable of supporting downstream applications like motion compensation, adaptive radiotherapy planning, and 4D image synthesis.

4.2.4. Ablation Study.

To better understand how different visual cues influence the zero-shot reasoning ability of the video model, we conduct an ablation study comparing three input settings: (i) *CT Only*, where the model learns directly from raw CT intensi-

Table 3. **Ablation study on input modalities.** We evaluate the impact of different input modalities on LVM performance using three configurations: (i) *CT Only*raw CT images without any segmentation guidance, (ii) *Mask Only*organ segmentation masks providing anatomical structure but no texture information, and (iii) *CT + Mask*a combined input that integrates both spatial texture and anatomical priors. Performance is assessed on lung and heart using a mixed dataset (20 public + 20 private 4D CT scans) and reported across four metrics: IoU, DSC, SD, and HD. The combination of CT and masks consistently yields the best performance across all organs and metrics, demonstrating that anatomical structure and image appearance provide complementary cues that jointly enhance motion modeling accuracy and boundary fidelity.

Data	Organ	Input	IoU (%) ↑	DSC (%) ↑	SD (mm) ↓	HD (mm) ↓
Mixed	Lung	CT Only	85.83	88.57	3.98	4.86
		Mask Only	87.60	91.88	3.19	4.29
		CT + Mask	91.23	94.87	2.75	4.18
	Heart	CT Only	83.71	86.83	4.11	5.93
		Mask Only	84.25	88.12	3.29	5.58
		CT + Mask	87.38	91.11	2.87	4.81
Private	Liver	CT Only	86.46	92.74	3.66	4.40
		Mask Only	88.64	93.97	3.22	4.12
		CT + Mask	92.95	96.35	2.95	3.98

Mixed: 20 4D CT scans of public dataset, 20 4D CT scans of our private dataset
IoU - intersection over union; DSC - dice similarity coefficient; SD - surface distance; HD - Hausdorff distance.

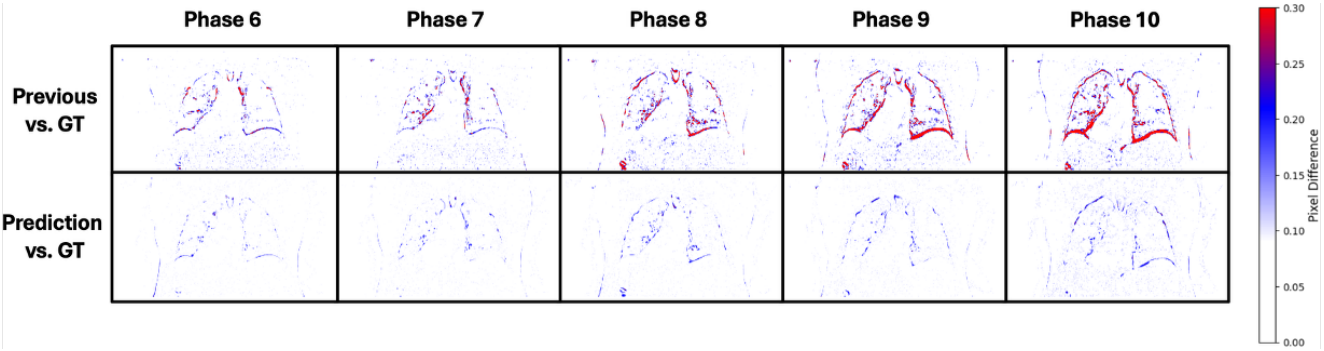

Figure 5. **Qualitative visualization of lung motion.** The first five phases are used as input, and the model predicts the next five. Each heatmap shows voxel-wise pixel differences between the ground truth (GT) and either the previous phase or the model prediction. Red indicates larger discrepancies. LVM accurately captures respiratory-induced motion, showing reduced errors and smoother temporal transitions compared to the prior phase, demonstrating coherent and anatomically consistent lung motion prediction.

ties; (ii) *Mask Only*, where the input is limited to segmentation masks highlighting organ structures; and (iii) *CT + Mask*, our default configuration combining both modalities. Results are summarized in Table 3 across lung, heart, and liver datasets.

With *CT Only*, the model leverages texture and intensity information to capture local deformation patterns, but lacks explicit structural guidance. This often results in blurred boundaries or overestimation of motion in background regions (e.g., 88.57% DSC for lung). When using *Mask Only*, the model focuses on organ contours and motion-specific regions, improving boundary alignment but losing fine-grained appearance cues essential for accurate deformation estimation. Combining both (*CT + Mask*) yields the best performance across all metrics (e.g., 94.87% DSC for lung, 96.35% for liver), demonstrating that CT intensities and segmentation masks provide complementary information: CT images encode texture and density continuity, while masks constrain motion learning to anatomically relevant regions.

These findings suggest that even in a zero-shot setting, large video models can effectively integrate multi-modal spatial cues to reason about organ structure and dynamics. The results highlight that combining pixel-level and semantic-level representations enhances both anatomical fidelity and temporal coherence, offering a more robust foundation for medical video understanding.

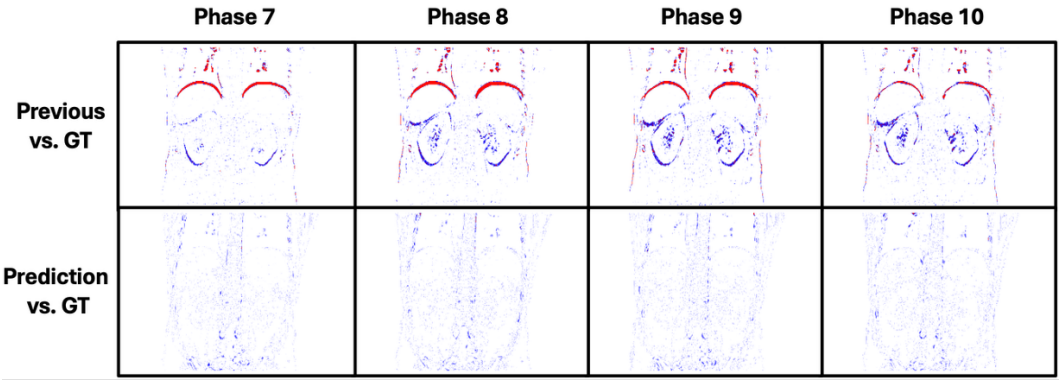

Figure 6. **Qualitative visualization of liver motion.** The first five 4D CT phases are used as input, and the model predicts the next five. Each heatmap shows voxel-wise differences between the prediction (or previous phase) and the ground truth, where red indicates larger errors. LVM accurately captures the livers smooth deformation and diaphragm-induced motion, maintaining temporal and anatomical consistency across phases.

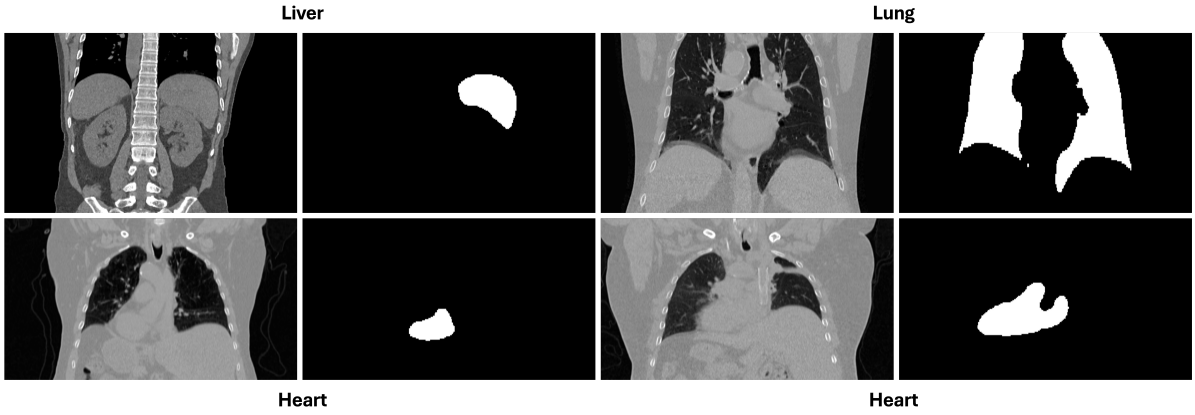

Figure 7. **Qualitative visualization of Segmentation.** For each organ, the left column shows the original CT slice, and the right column shows the predicted segmentation mask. The results demonstrate that the zero-shot video model can accurately segment organs across diverse anatomical regions based on the given input prompts.

4.2.5. Visualization

To qualitatively evaluate the effectiveness of LVM in modeling organ motion, we present representative visualizations of predicted versus ground-truth CT phases for different anatomical regions. While quantitative metrics such as DSC and HD provide overall accuracy, qualitative visualization enables intuitive assessment of temporal consistency, local deformation patterns, and anatomical boundary fidelityfactors critical for clinical interpretability and radiotherapy planning. We highlight examples for lung and liver 4D CT sequences, demonstrating LVM's capacity to generate smooth, realistic motion pattern aligned with patient-specific breathing dynamics.

Lung 4D CT Motion Prediction. Figure 5 presents a qualitative heatmap analysis of lung motion prediction for a representative thoracic radiotherapy case. The first five 4D CT phases are used as input, and the model predicts the next five phases. Each heatmap shows voxel-wise pixel differences between the predicted or previous phase and the ground truth (GT), with red indicating larger discrepancies. Compared to the prior phase, the prediction from LVM demonstrates significantly reduced errors across lung boundaries and internal structures, indicating accurate modeling of respiratory-induced expansion and contraction. The results highlight LVMs ability to learn patient-specific breathing dynamics and maintain temporal and anatomical consistency across multiple phaseskey requirements for motion-adaptive radiotherapy planning.

Liver 4D CT Motion Prediction. Figure 6 shows a qualitative heatmap analysis of liver motion prediction under respiratory influence. Similar to the lung case, the model is provided with the first five 4D CT phases and tasked to predict the next five. Each heatmap represents voxel-wise differences between the predicted or previous phase and the

ground truth (GT), where red indicates higher pixel deviation. LVM effectively captures the smooth deformation and positional shift of the liver caused by diaphragm motion, showing substantially reduced discrepancies compared to the prior phase. This demonstrates the models capability to reason about complex abdominal motion while maintaining anatomical fidelity over time critical requirement for accurate dose planning and delivery in liver radiotherapy.

Segmentation. Figure 7 shows qualitative examples of zero-shot organ segmentation for the liver, heart, and lung. Without any task-specific fine-tuning, the video model accurately identifies anatomical structures and delineates organ boundaries directly from CT images. These results suggest that the model can generalize to spatial reasoning tasks in medical imaging, effectively transferring its visual understanding from natural scenes to clinical contexts.

These qualitative visualizations highlight the potential of the model to advance motion management by providing predictions with patient-specific motion patterns, which are critical for adaptive treatment planning.

5. Limitations & Future Directions

Although our study provides the first evidence that zero-shot video models can perform clinically meaningful reasoning on medical imaging data, several challenges remain before such models can be reliably integrated into clinical radiotherapy workflows.

Beyond intra-fraction motion. Our current work primarily focuses on intra-fraction motion prediction, capturing respiratory-induced organ motion within a single 4D CT session. However, real-world radiotherapy treatments span multiple weeks, during which inter-fraction anatomical changes such as tumor shrinkage, weight loss, and baseline organ shifts frequently occur. Modeling these long-term variations will require incorporating longitudinal imaging across multiple sessions and reasoning over extended temporal contexts. Extending video models to handle such inter-fraction dynamics represents an important direction for future research.

Generalization across fundamental imaging tasks. Our exploration of segmentation, denoising, and super-resolution tasks demonstrates that large video models can already generalize across fundamentally different medical imaging problems in a zero-shot setting. Although the current performance does not yet match specialized, domain-tuned networks, the results reveal key signs of transferable spatial reasoning and cross-task adaptability. The model shows an ability to localize anatomical structures, preserve fine textures, and recover structural continuitywithout any medical-domain training. These findings highlight that even without explicit optimization, video models inherently learn temporal and structural priors that align with medical imaging needs, indicating their strong potential as a universal visual backbone for clinical applications.

Toward a unified medical foundation model. Ultimately, these efforts aim toward a unified medical foundation model built upon large video architecturescapable of performing segmentation, enhancement, motion prediction, and temporal reasoning across diverse clinical tasks and modalities, all without task-specific retraining. Such a model would represent a transformative step toward general-purpose, adaptive medical AI that can learn, reason, and generalize across space and time.

6. Conclusion

In this work, we explore the zero-shot capabilities of large video models in the medical imaging domain. Using radiotherapy organ motion prediction as a challenging testbed, we show that a general-purpose video modelwithout any medical-domain trainingcan learn and reason about patient-specific respiratory motion directly from sequential CT data. The model produces temporally coherent and anatomically consistent predictions, surpassing specialized pipelines that rely on deformation-field supervision or task-specific architectures.

Beyond motion prediction, we demonstrate that the same model can perform segmentation, denoising, and super-resolution in a zero-shot manner. Although its performance does not yet match fully supervised approaches, these results clearly reveal inherent adaptability and cross-task generalization of large video models. Such capabilities mark significant step toward unified visual model in medicine.

Looking forward, we envision extending this line of research toward developing medical foundation models built upon large video architecturescapable of performing diverse imaging tasks with minimal supervision. This paradigm promises scalable, adaptive, and clinically relevant AI systems that can learn, interpret, and reason across time, anatomy, and modality.

Acknowledgments

This research is supported in part by the National Institutes of Health under Award Numbers R01EB032680, R01CA272991, R01DE033512, and U54CA274513.

References

[1] Yutong Bai, Xinyang Geng, Karttikeya Mangalam, Amir Bar, Alan L Yuille, Trevor Darrell, Jitendra Malik, and Alexei A Efros. Sequential modeling enables scalable learning for large vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22861–22872, 2024.

[2] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[5] Qi Chen, Yuxiang Lai, Xiaoxi Chen, Qixin Hu, Alan Yuille, and Zongwei Zhou. Analyzing tumors by synthesis. *Generative Machine Learning Models in Medical Image Computing*, page 85, 2024.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2020.

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[8] Zahra Ghasemi and Payam Samadi Miandoab. Feasibility study of convolutional long shortterm memory network for pulmonary movement prediction in ct images. *Journal of Biomedical Physics & Engineering*, 14(1):55, 2024.

[9] Google. Veo 3 announcement. https://blog.google/technology/ai/generative-media-models-io-2025/, 2025. Accessed: September 22, 2025.

[10] Google. Veo 3 launch. https://cloud.google.com/blog/products/ai-machine-learning/veo-3-fast-available-for-everyone-on-vertex-ai, 2025. Accessed: September 22, 2025.

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[12] Geoffrey D Hugo, Elisabeth Weiss, William C Sleeman, Salim Balik, Paul J Keall, Jun Lu, and Jeffrey F Williamson. A longitudinal four-dimensional computed tomography and cone beam computed tomography dataset for image-guided radiation therapy research in lung cancer. *Medical physics*, 44(2):762–771, 2017.

[13] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

[14] Yune Kwong, Alexandra Olimpia Mel, Greg Wheeler, and John M Troupis. F our-dimensional computed tomography (4dct): a review of the current status and applications. *Journal of medical imaging and radiation oncology*, 59(5):545–554, 2015.

[15] Yuxiang Lai, Yi Zhou, Xinghong Liu, and Tao Zhou. Memory-assisted sub-prototype mining for universal domain adaptation. *arXiv preprint arXiv:2310.05453*, 2023.

[16] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 36–46. Springer, 2024.

[17] Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025.

[18] Donghoon Lee, Ellen Yorke, Masoud Zarepisheh, Saad Nadeem, and Yu-Chi Hu. Rmsim: controlled respiratory motion simulation on static patient scans. *Physics in Medicine & Biology*, 68(4):045009, 2023.

[19] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, 97:103285, 2024.

[20] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023.

[21] Ke Nie, Cynthia Chuang, Neil Kirby, Steve Braunstein, and Jean Pouliot. Site-specific deformable imaging registration algorithm selection using patient-based simulated deformations. *Medical physics*, 40(4):041911, 2013.

[22] OpenAI. Sora 2 system card. https://openai.com/index/sora-2-system-card/, 2025. Accessed: September 22, 2025.

[23] Oscar Pastor-Serrano, Steven Habraken, Mischa Hoogeman, Danny Lathouwers, Dennis Schaart, Yusuke Nomura, Lei Xing, and Zoltán Perkó. A probabilistic deep learning model of inter-fraction anatomical variations in radiotherapy. *Physics in Medicine & Biology*, 68(8):085018, 2023.

[24] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[25] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image

synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[26] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[27] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

[28] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[29] Andreas Smolders, Luciano Rivetti, Nadine Vatterodt, Stine Korreman, Anthony Lomax, Manju Sharma, Andrej Studen, Damien Charles Weber, Robert Jeraj, and Francesca Albetini. Diffusert: predicting likely anatomical deformations of patients undergoing radiotherapy. *Physics in Medicine & Biology*, 69(15):155016, 2024.

[30] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[31] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[33] Douglas J Vile. *Statistical modeling of interfractional tissue deformation and its application in radiation therapy planning*. Virginia Commonwealth University, 2015.

[34] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, page 3876, 2022.

[35] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5), 2023.

[36] Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[37] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[38] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.