This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# EchoWorld: Learning Motion-Aware World Models for Echocardiography Probe Guidance

Yang Yue[1*] Yulin Wang[1*] Haojun Jiang[1] Pan Liu[2] Shiji Song[1] Gao Huang[1✉]
[1]Tsinghua University [2]PLA General Hospital
yueyang22@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn

## Abstract

*Echocardiography is crucial for cardiovascular disease detection but relies heavily on experienced sonographers. Echocardiography probe guidance systems, which provide real-time movement instructions for acquiring standard plane images, offer a promising solution for AI-assisted or fully autonomous scanning. However, developing effective machine learning models for this task remains challenging, as they must grasp heart anatomy and the intricate interplay between probe motion and visual signals. To address this, we present EchoWorld, a motion-aware world modeling framework for probe guidance that encodes anatomical knowledge and motion-induced visual dynamics, while effectively leveraging past visual-motion sequences to enhance guidance precision. EchoWorld employs a pre-training strategy inspired by world modeling principles, where the model predicts masked anatomical regions and simulates the visual outcomes of probe adjustments. Built upon this pre-trained model, we introduce a motion-aware attention mechanism in the fine-tuning stage that effectively integrates historical visual-motion data, enabling precise and adaptive probe guidance. Trained on more than one million ultrasound images from over 200 routine scans, EchoWorld effectively captures key echocardiographic knowledge, as validated by qualitative analysis. Moreover, our method significantly reduces guidance errors compared to existing visual backbones and guidance frameworks, excelling in both single-frame and sequential evaluation protocols. Code is available at* https://github.com/LeapLabTHU/EchoWorld.

## 1. Introduction

Cardiovascular disease remains one of the leading causes of death worldwide [46, 52], making timely and accurate diagnosis critical to saving lives. Among the various diagnostic tools available, echocardiography stands out as a non-invasive, cost-effective, and widely accessible method for assessing cardiac health. In this method, a *probe* emits high-frequency sound waves into the body, which are reflected by heart structures and captured to generate real-time images (Figure 1(a)). However, performing cardiac ultrasound scans requires the sonographer to carefully maneuver the probe to acquire key sectional views of the heart (Figure 1(a-b)), a task that demands extensive anatomical knowledge and experience. This complexity, coupled with a global shortage of qualified sonographers, limits the accessibility of ultrasound services, particularly in less developed regions. This challenge has motivated the development of probe guidance systems [19, 29, 40, 48] capable of assisting less experienced sonographers or, in the longer term, enabling fully autonomous ultrasound scanning robots. Such systems hold the potential to democratize cardiac care by providing real-time, actionable feedback, which can significantly improve the efficiency of the scanning process.

As shown in Figure 1(c), probe guidance in echocardiography can be formulated as a vision-based sequential prediction problem, where a model needs to predict the necessary probe movement vectors to reach each target view, utilizing historical visual-motion data. However, unlike other computer vision tasks in medical imaging, it presents the unique challenge of integrating motion data with dynamic visual observations. Autonomous systems must not only understand the complex anatomical structures of the heart (*e.g.* chambers, valves, and vessels) but also how these structures are represented in ultrasound images as the probe moves and changes position. While early efforts in probe guidance [19, 40, 50] have made some progress in developing assistive and autonomous scanning systems, few studies focus on a fundamental problem: *How can we develop a principled approach that effectively learns essential medical knowledge while seamlessly integrating visual and motion data for precise probe guidance?*

In this paper, we present EchoWorld, a motion-aware world modeling framework that begins by pre-training a

*Equal contribution. ✉Corresponding authors.

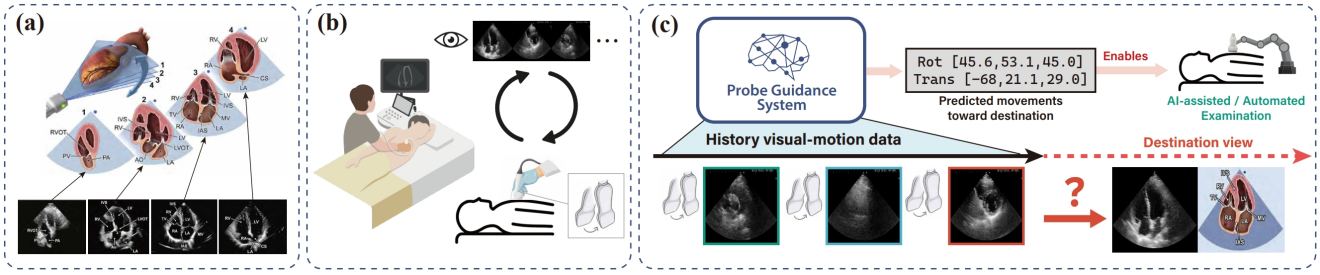


Figure 1. **Overview of cardiac ultrasound and the probe guidance task.** (a) The ultrasound probe captures cross-sectional views of the heart, with variations in probe position and orientation corresponding to different anatomical structures. (b) During the ultrasound scanning process, the sonographer maneuvers the probe on the patient's chest, continuously adjusting its position and orientation based on real-time visual feedback. (c) A probe guidance system can potentially automate the scanning process by predicting the necessary probe movements to reach a target view, utilizing historical visual-motion data.

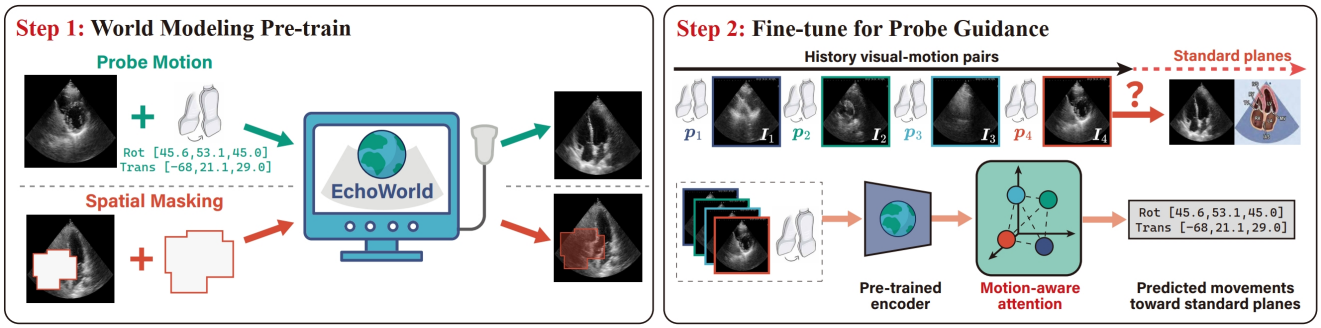


Figure 2. **Overview of the proposed framework.** Left: We pre-train a cardiac world model to capture ultrasound knowledge through spatial and motion modeling tasks. Right: The pre-trained model is fine-tuned for probe guidance, incorporating a motion-aware attention mechanism to effectively integrate visual-motion features.

strong representation model on visual-motion data, followed by fine-tuning the model with a novel motion-aware attention mechanism that allows seamless integration of motion information with visual features.

The first stage, world model pre-training, is designed to encode rich, common-sense knowledge about the world [23, 34], which can potentially capture the heart's anatomical structure and the relationships between different probe positions, allowing the system to guide a sonographer much like an experienced driver navigating through the city with an internalized map. As demonstrated in the left part of Figure 2, our cardiac world model encodes two key dimensions of echocardiology knowledge: 1) the appearance of anatomical structures (*e.g.*, ventricle, valves, and septums) in cardiac ultrasound images and 2) the changing dynamic of visual signals following the probe motions.

Building on this pre-trained world model, we introduce a guidance prediction module with a motion-aware attention mechanism that integrates historical image-pose data, as shown in the right part of Figure 2. Unlike existing methods that typically organize the data into interleaved visual-action sequences, our motion-aware attention mechanism embeds 3D relative pose differences into the attention features. This enables motion-aware interactions across image frames, allowing the system to better track anatomical structures and produce more accurate predictions.

Our proposed framework, EchoWorld, is built upon a cardiac ultrasound scanning dataset derived from routine clinical examinations. We empirically show that EchoWorld can act as a cardiac ultrasound simulator enriched with anatomical knowledge. Additionally, we compare EchoWorld against a wide range of pre-trained models and existing probe guidance methods. Our model consistently outperforms these approaches in acquiring ten standard planes, achieving lower guidance errors across both single-frame and sequential evaluation protocols. Analytical results and ablation studies further validate the effectiveness of the proposed framework.

## 2. Related Work

**World Models.** The concept of the world model was first introduced in psychology [16] and later adapted for model-predictive control [10, 11] and reinforcement learning [22–25]. In these contexts, a world model typically predicts future states of the environment based on an agent's actions. More recently, the development of general-purpose world models that encompass broad, commonsense understanding has been recognized as a key step toward achieving general artificial intelligence [34, 47]. The advent of large-scale video generation models has further highlighted this potential, demonstrating their ability to serve as physical

simulators [32, 43], driving scene simulators [20, 28, 62], and game engines [1, 2, 57]. World modeling has also proven to be a powerful tool for representation learning, producing structured and informative representations that capture complex world dynamics with minimal supervision [4–7, 21, 24–26]. Our proposed framework utilizes world modeling as a pretext task to develop a robust visual representation model. Our analytical experiments show that EchoWorld effectively captures echocardiography knowledge for probe guidance. When augmented with a diffusion model, it can potentially function as a simulator for free-hand cardiac ultrasound scanning, capable of predicting visual changes based on probe movements.

**AI for ultrasound.** Recent AI advances have driven significant progress in ultrasound applications, advancing tasks like segmentation [17, 36, 54], 3D reconstruction [61], and diagnostic support [45, 59]. Recent efforts also focus on creating foundation models for ultrasound, as exemplified by USFM [31], which explores self-supervised learning, and EchoCLIP [15], which utilizes multimodal learning techniques. Another critical AI application in ultrasound is probe guidance, aimed at assisting novices and inexperienced sonographers [19, 40, 48] or enabling fully autonomous robotic scanning [50, 51]. For instance, US-GuideNet [19] provides rotational guidance for free-hand obstetric ultrasound, while Shida et al. [50, 51] investigates search-based methods to acquire the Parasternal Long-Axis (PLAX) plane in cardiac ultrasound. Current ultrasound probe guidance methods primarily leverage imitation learning [19, 29, 30, 37] or reinforcement learning [3, 35]. The latter typically relies on CT-derived simulations , whereas imitation learning—the approach adopted in our study—directly learns from expert demonstrations, presenting a scalable approach [29] in line with advances in general-purpose robotic control [9, 33, 41]. While prior research has focused on probe control for ultrasound scanning, little attention has been given to representation learning strategies and network architectures for ultrasound data. In this paper, we seek to bridge these gaps by proposing a motion-aware world modeling framework tailored for ultrasound.

## 3. Background and Notations

Before presenting our method, we briefly overview the technical background relevant to the probe guidance task and the format of the data employed in our study.

### 3.1. Cardiac Ultrasound

Cardiac ultrasound, or echocardiography, involves the use of a transducer (or *probe*) that emits high-frequency sound waves into the body, which are then reflected by heart structures and captured to generate real-time images. These images depict two-dimensional cross-sectional views, or *planes*, of the heart, showcasing its chambers, walls, valves, and blood flow dynamics, and are essential for diagnosing a variety of cardiac conditions. The specific position, orientation, and tilt of the probe determine the captured plane, as shown in Figure 1(a). A *standard plane* in cardiac ultrasound is a predefined cross-sectional view, systematically used to assess specific heart structures. For instance, the Parasternal Short-Axis (PSAX) plane offers a horizontal slice through the heart, which is instrumental in evaluating heart muscle thickness and valve function. Obtaining these views requires precise probe maneuvers; the sonographer skillfully maneuvers the probe on the patient's chest, carefully adjusting its position, angle, and pressure to capture the desired two-dimensional "slice" of the heart.

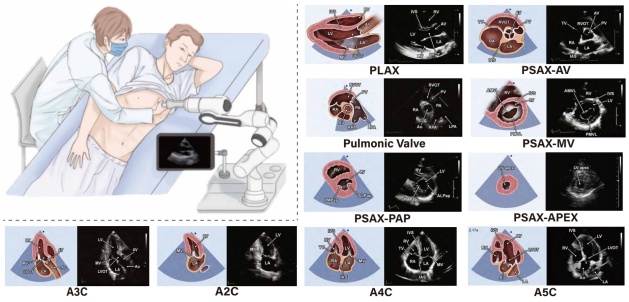


Figure 3. Illustration of our dataset and task. Top-left: We collect expert demonstration data where the sonographer controls a robot arm with a probe, recording both image frames and probe motion synchronously. Remaining figure: The ten standard planes targeted for acquisition. Figures are adapted from [29, 39].

During a typical echocardiographic examination, the sonographer sequentially captures several standard planes, measuring parameters in each view before forwarding the data for cardiologist interpretation. The probe guidance task, aimed at supporting or potentially automating this process, involves developing models that can direct the probe towards specified target planes. This automation could streamline the acquisition of accurate and diagnostically relevant images, making echocardiography more accessible and consistent across varied operator skill levels.

### 3.2. Dataset and Task

Our study is conducted based on an expert demonstration dataset collected during routine clinical ultrasound exams, in which professional sonographers maneuvered an ultrasound probe mounted on a robotic arm. This setup enables synchronous recording of both image frames and probe pose information, as illustrated in Figure 3. A detailed introduction of the dataset can be found in the appendix.

The dataset is organized by "*scans*", each representing a recorded examination of a patient. A scan comprises a multi-minute echocardiography video (30 fps) and the probe's corresponding pose for each frame in the anatomical coordinate system. Formally, each scan forms a visual-motion sequence $\{(\boldsymbol{I}_t, \boldsymbol{p}_t)\}_{t=1}^{T}$, where $T$ is the total number

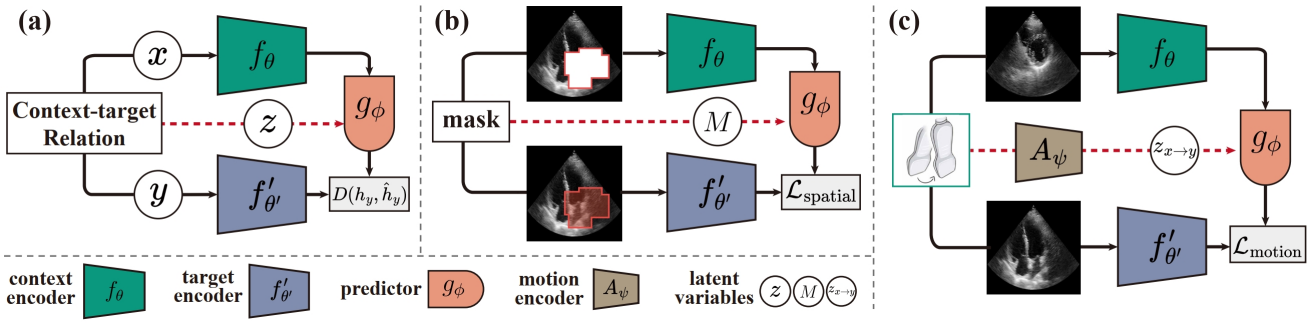


Figure 4. **Illustration of the world modeling tasks.** (a) A basic world modeling framework [34], where the task is to predict the target $y$ from context $x$ in feature space, using a latent variable $z$ encoding their relationship. (b) The spatial world modeling task, which recovers masked anatomical structures. (c) The motion world modeling task, which predicts visual changes in the context based on probe motion.

of timesteps, $\boldsymbol{I}_t$ is the ultrasound image at time $t$, and $\boldsymbol{p}_t$ denotes the probe pose. The probe poses are represented in six degrees of freedom (6-DOF): three translational coordinates (x, y, z) and three rotational components (yaw, pitch, roll) in Euler angles. Each probe pose belongs to the rigid transformation group, allowing us to express the *relative movement* between two poses $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ as $\boldsymbol{p}_{j\to i} = \boldsymbol{p}_i \cdot \boldsymbol{p}_j^{-1}$, where "$\cdot$" denotes the composition of transformations. While absolute probe poses can vary substantially across scans, the relative movements exhibit consistent patterns that reveal the sonographer's probe maneuvers.

During each scan, the sonographer visits multiple standard planes, with corresponding timestamps and plane types labeled as ground truth for the probe guidance task. Our goal is to predict the relative movement from each frame in the dataset to these target plane poses. Suppose the current timestep is $t_0$ and the target standard plane is reached at probe pose $\boldsymbol{p}^*$, then the ground truth movement at $t_0$ is given by $\boldsymbol{a}_{t_0} = \boldsymbol{p}^* \cdot \boldsymbol{p}_{t_0}^{-1}$. The probe guidance model leverages historical visual-motion data, $\{(\boldsymbol{I}_t, \boldsymbol{p}_t)\}_{t\le t_0}$, to predict the movement $\hat{\boldsymbol{a}}_{t_0}$ needed to reach the target plane. We focus on ten standard planes based on their clinical importance and prevalence in the dataset, as shown in Figure 3.

## 4. EchoWorld

This section presents our motion-aware world modeling framework for probe guidance, with the overall concept illustrated in Figure 2. We adopt a two-stage approach: a pre-training phase to construct a cardiac world model that captures essential cardiac ultrasound knowledge, followed by a fine-tuning phase with a motion-aware attention mechanism for integrating historical visual-motion data effectively.

### 4.1. Pre-training Cardiac World Models

Humans are believed to maintain internal models that capture complex patterns and dynamics of the world [16, 23, 34], encoding prior knowledge that supports perception, planning, and decision-making. Similarly, an experienced sonographer develops a mental model of the heart's structures and can anticipate visual changes as they adjust the probe. Inspired by this, we design world modeling tasks that equip our model with a similar understanding of cardiac anatomy and motion.

**A basic world modeling framework.** Our world model is based on the joint-embedding predictive architecture (JEPA) [34], illustrated in Figure 4(a). The model's objective is to predict a target $y$, an unobserved portion of the world (such as the ultrasound scan in our study), using the available context $x$, conditioned on a latent variable $z$ that captures the relationship between the contexts and targets (*e.g.* the probe movement that leads to the visual changes).

The JEPA consists of a *context encoder* $f_\theta$ and a *target encoder* $f'_{\theta'}$, which produce context and target features, $h_x = f_\theta(x)$ and $h_y = f'_{\theta'}(y)$. A *predictor* $g_\phi$ then predicts the target features conditioned on the latent variable $z$:

$$\hat{h}(y) = g_\phi(h_x, z) = g_\phi(f_\theta(x), z). \tag{1}$$

The latent variable $z$ encodes essential information that bridges the context $x$ and target $y$, capturing the uncertainty and inherent dynamics of the world. The model's objective is to minimize prediction error:

$$\text{minimize} \quad D(h_y, \hat{h}_y) = D(f'_{\theta'}(y), g_\phi(f_\theta(x); z)), \tag{2}$$

where $D$ is an error function. To prevent feature collapse, the target encoder $f'_{\theta'}$ is an exponential moving average (EMA) of the context encoder. Below, we detail two world modeling tasks tailored for cardiac ultrasound.

**Spatial world modeling.** Ultrasound images reveal a complex array of anatomical structures within the heart, including chambers, valves, and arteries, as shown in Figure 3. A thorough understanding of these structures' layout and visual characteristics is essential for sonographers. To instill this anatomical knowledge in our model, we incorporate a mask-and-reconstruction task [4, 5, 27, 60], in which several contiguous regions of the image are masked, and the model is tasked with predicting the features for these areas. This task encourages the model to learn localized textures and the spatial relationships among anatomical structures.

Our model uses vision transformers as the context and target encoders, with masking achieved by dropping rectangular blocks of visual patches (Figure 4(b)). Only visible patches are passed to the context encoder, while the target encoder processes the entire image:

$$h_x = f_\theta(\text{Mask}(\boldsymbol{I})), \quad h_y = f'_{\theta'}(\boldsymbol{I}). \tag{3}$$

In the predictor, mask tokens $m$ carry the encoded positions of the masked regions, serving as the latent variable $z$. Letting $M$ denote the masked patch locations, the predictor output $\hat{h}_y$ is given by:

$$\hat{h}_y = g_\phi\left((h_x + p_x) \oplus \{m + \text{PE}(c)\}_{c \in M}\right), \tag{4}$$

where $\oplus$ denotes concatenation, $p_x$ is the positional embedding of the context tokens, $c \in M$ are 2D patch coordinates, and $\text{PE}(\cdot)$ represents sinusoidal positional encoding [58]. The loss is computed only over the masked locations:

$$\mathcal{L}_{\text{spatial}} = \sum_{c \in M} \|g_\phi(f_\theta(x); M)_c - f'_{\theta'}(y)_c\|_1 . \tag{5}$$

**Motion world modeling.** Beyond anatomical understanding in the 2D image plane, grasping the dynamics of ultrasound scanning is crucial for accurate probe guidance. As the probe is tilted, rotated, or moved, the visual features of cardiac structures—such as chamber alignment, valve orientation, and surrounding tissue—shift accordingly. Motion world modeling aims to capture the relationship between probe movements and these resulting visual changes. To build a motion-aware model, we introduce a predictive task in which the model anticipates changes in visual features based on probe movements. Specifically, given two images $\boldsymbol{I}_a, \boldsymbol{I}_b$ (serve as context and target) from the same scan, along with their relative probe movement $\boldsymbol{p}_{a\to b}$, we encode the motion information using a motion encoder $A_\psi$, yielding motion features $\boldsymbol{z}_{a\to b} = A_\psi(\boldsymbol{p}_{a\to b})$. A single mask token carries these motion features, which are fed into the predictor along with the context features to predict the average-pooled target features:

$$\begin{aligned} \hat{h}_y &= g_\phi(f_\theta(\boldsymbol{I}_a); m + \boldsymbol{z}_{a\to b}), \\ h_y &= \text{AvgPool}(f'_{\theta'}(\boldsymbol{I}_b)). \end{aligned} \tag{6}$$

To enhance the model's sensitivity to probe motion, we employ a contrastive objective [53], where the prediction $\hat{h}_y$ and the corresponding target $h_y$ serve as positive pairs. Following [13], both $\hat{h}_y, h_y$ are projected through a two-layer MLP[1] before applying the InfoNCE loss [42]:

$$\mathcal{L}_{motion} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\hat{h}_{y_i}^\top \cdot h_{y_i}/\tau)}{\sum_j \exp(\hat{h}_{y_i}^\top \cdot h_{y_j}/\tau)}, \tag{7}$$

where $B$ is the batch size.

**Joint Modeling.** EchoWorld integrates spatial and motion world modeling into a unified pre-training approach designed to maximize the complementary strengths of each task. Given two images $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$ from the same scan, spatial world modeling is applied independently to both images, while motion world modeling uses $\boldsymbol{p}_{a\to b}$ to learn the transition between them. The final loss function is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spatial}} + \lambda \mathcal{L}_{\text{motion}}, \tag{8}$$

where $\lambda = 0.1$ balances the scale of the two losses. Combining these two modeling strategies creates a holistic representation where each image is understood both in its anatomical context and as part of a dynamic process.

[1]The projector is omitted in Eq (7) for brevity.

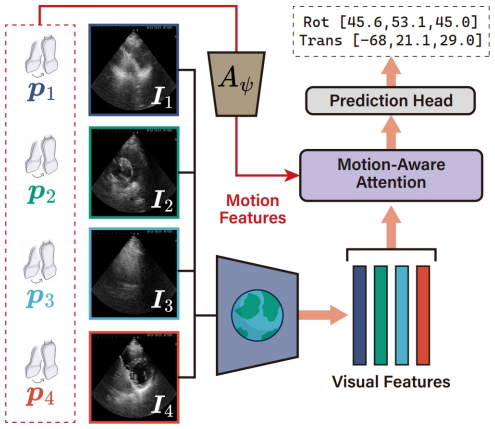


Figure 5. **The probe guidance pipeline.** Given a sequence of historical visual-motion pairs, we first extract features using the pre-trained visual and motion encoders. These features are then integrated via a motion-aware attention mechanism and projected to the final guidance output.

## 4.2. Motion-Aware Probe Guidance

With the pre-trained world model in place, equipped with representations that capture both cardiac anatomy and visual-motion dynamics, we can now introduce a guidance predictor tailored for probe guidance applications. As shown in Figure 5, we introduce a motion-aware attention module that integrates historical visual-motion data by embedding motion information into attention features, enabling more effective aggregation of past observations.

Specifically, the probe guidance model predicts the probe movements to reach target planes by processing a sequence of visual-motion signal $\{I_{t_i}, p_{t_i}\}_{i=1}^N$, where $N$ is the sequence length, $t_1 \le t_2 \le \cdots \le t_N$ with $t_N$ representing the latest timestep[2]. Previous methods typically organize historical data into an interleaved image-action sequence $\{\boldsymbol{I}_{t_1}, \boldsymbol{p}_{t_1\to t_2}, \boldsymbol{I}_{t_2}, \boldsymbol{p}_{t_2\to t_3}, \boldsymbol{I}_{t_3}, \ldots, \boldsymbol{I}_{t_N}\}$ and encode the sequence using a recurrent model. However, we argue that this formulation is suboptimal as it fails to fully utilize the rich motion data available. Instead, we propose a motion-aware attention mechanism that incorporates motion signals

[2]Since the original videos contain thousands of images, we sample a subset of timesteps as input to the model.

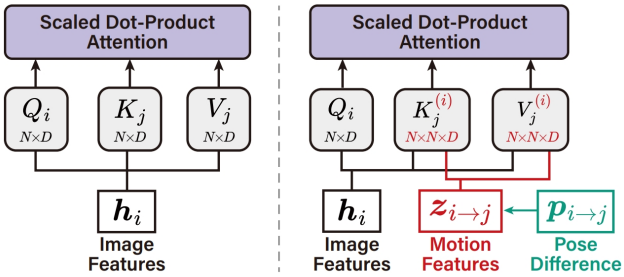


Figure 6. **Comparison of standard (left) and motion-aware (right) attention,** where the latter incorporates pairwise pose differences across the tokens into the key-value pairs.

directly into the interactions between all frame tokens. This approach aligns the attention mechanism with the 3D spatial transformations of the probe, facilitating more precise modeling of the spatial relationships between frames.

Specifically, we first use the pre-trained world model to extract the features of the visual and motion data:

$$\boldsymbol{h}_i = \text{AvgPool}(f_\theta(\boldsymbol{I}_{t_i})), \quad \boldsymbol{z}_{i\to j} = A_\psi(\boldsymbol{p}_{t_i\to t_j}). \tag{9}$$

The average-pooled visual features $\{\boldsymbol{h}_i\}_{i=1}^N$ serve as input tokens to our motion-aware attention module. The standard scaled dot-product attention is computed as follows:

$$O_i = \sum_j \frac{\exp(Q_i^\top K_j)}{\sum_l \exp(Q_i^\top K_l)} V_j, \tag{10}$$

where $O_i$ is the output representation of token $i$. Since this operation is permutation invariant, position information must be embedded in the tokens for the model to understand spatial relationships. In general transformer architectures, this is achieved with absolute [58] or relative [49, 55] positional embeddings. For probe guidance, however, it is essential to embed motion-specific information into the attention mechanism to capture the relative spatial transformations of the probe. We achieve this by encoding the pairwise pose differences into the key and value features:

$$K_j^{(i)} = \text{MLP}(\boldsymbol{h}_j, \boldsymbol{z}_{i\to j}),\ V_j^{(i)} = \text{MLP}(\boldsymbol{h}_j, \boldsymbol{z}_{i\to j}). \tag{11}$$

This allows each query token to be associated with a unique set of key-value pairs that encode the relative pose information. The attention output for token $i$ is computed as:

$$O_i = \sum_j \frac{\exp(Q_i^\top K_j^{(i)})}{\sum_l \exp(Q_i^\top K_l^{(i)})} V_j^{(i)}. \tag{12}$$

This motion-aware approach contrasts with standard self-attention by incorporating relative motion information, enabling better capture of spatial changes induced by probe movements. Additionally, we employ a multi-head design [58] to capture diverse attention patterns. Figure 6 compares the standard and motion-aware attention mechanisms.

After the motion-aware attention, the aggregated features are fed into a prediction head to regress the probe movement to the target plane. The guidance loss is computed as $\mathcal{L}_{\text{guide}} = \|\boldsymbol{a}_t - \hat{\boldsymbol{a}}_t\|_1$, where $\hat{\boldsymbol{a}}_t$ is the model's predicted movement to the target plane at timestep $t$ and $\boldsymbol{a}_t$ is the ground truth movement.

## 5. Experiments

In this section, we first demonstrate EchoWorld's world modeling capabilities via visualizations. We then compare EchoWorld with existing methods on the probe guidance task using two evaluation protocols: single-frame and sequential. Finally, ablation studies highlight the contribution of each framework component.

**Implementation details.** Our dataset comprises 356 routine cardiac ultrasound scans, collected and processed as described in Section 3.2. The dataset consists of approximately one million image frames, each paired with its corresponding probe pose. The data is divided into 284 scans for training and 72 scans for testing, with no overlap in patients between the two splits. For pre-training, we employ a ViT-Small [58] as the visual backbone and use a transformer with six blocks as the predictor. Target encoder $f'_{\theta'}$ is the exponential moving average of the context encoder $f_\theta$. Training is conducted with a batch size of 1024 over 300 epochs on four A100 GPUs. After pre-training, the model is fine-tuned using ground truth movements toward the standard planes.

### 5.1. EchoWorld as a World Model

To validate the world modeling capabilities of EchoWorld, we train a generative model that maps the predicted features $\hat{h}_y$ back to pixel space using the RCDM framework [8]. The context-target pairs are constructed in the same manner as in the spatial and motion world modeling tasks. A diffusion model is then trained to generate the target image conditioned on the predicted features $\hat{h}_y$ (average pooled for spatial modeling). As shown in Figure 7(a), EchoWorld successfully recovers masked regions that are consistent with the context, demonstrating its understanding of the appearance and spatial relationship between anatomical structures. In Figure 7(b), EchoWorld simulates ultrasound scanning by predicting visual changes based on probe movements. This ability highlights the model's capacity to capture the interaction between visual and motion signals, which is crucial for guiding the probe toward the destination.

### 5.2. EchoWorld for Probe Guidance

**Evaluation protocol.** Evaluating different approaches in real-person experiments is both costly and time-consuming. To address this, we propose two evaluation protocols based on the collected scanning data, as detailed below:

*Single-frame protocol*: In this protocol, a two-layer MLP head is attached to each visual backbone to predict probe movements to ten target planes from a single ultrasound image. This task focuses on evaluating the representation

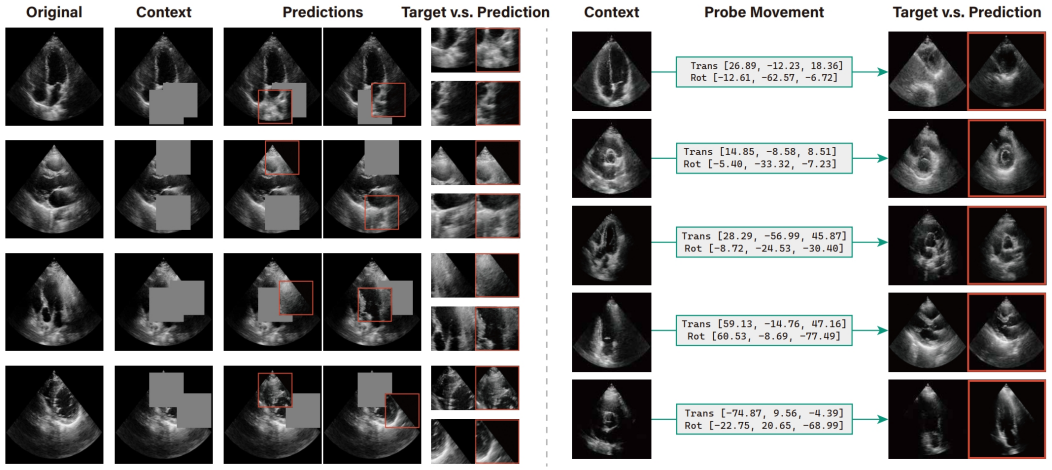


Figure 7. **Visualization of world modeling predictions.** Using a diffusion model, we map the outputs of the predictor to pixel space. The predictor effectively recovers (a) masked anatomical regions and (b) the visual changes resulting from probe movement. The predicted regions are highlighted with red bounding boxes.

Table 1. **Performance comparison of EchoWorld and baselines on the probe guidance task**. Experiments are conducted under both the single-frame and sequential protocols. The mean absolute error is reported. "Trans." denotes the translation error (x, y, z) in millimeters, while "Rot." represents the rotation error (yaw, pitch, roll) in degrees. The best two results are **bold-faced** and underlined, respectively.

| Method | PLAX | | PSAX-AV | | PSAX-PV | | PSAX-MV | | PSAX-PAP | | PSAX-APEX | | A4C | | A5C | | A3C | | A2C | | Avg |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | Trans. | Rot. | |
| *Single-frame protocol, not pre-trained on ultrasound data* | | | | | | | | | | | | | | | | | | | | | |
| Scratch [18] | 8.84 | 7.78 | 8.12 | 8.31 | 8.65 | 9.02 | 8.18 | 9.21 | 8.05 | 9.13 | 9.56 | 9.89 | 8.86 | 7.95 | 9.06 | 10.04 | 9.35 | 10.22 | 9.09 | 11.97 | 9.07 |
| DeiT [56] | 8.50 | 7.27 | 8.09 | 8.02 | 8.51 | 8.71 | 7.90 | 8.66 | 7.85 | 8.62 | 9.21 | 9.23 | 8.43 | 7.37 | 8.55 | 9.54 | 8.70 | 9.71 | 8.52 | 11.28 | 8.63 |
| DINOv2 [44] | 8.31 | 7.21 | 7.82 | 7.82 | 8.36 | 8.75 | 7.70 | 8.66 | 7.66 | 8.42 | 9.09 | 9.04 | 8.40 | 7.27 | 8.52 | 9.41 | 8.70 | 9.38 | 8.62 | 11.27 | 8.52 |
| *Single-frame protocol, pre-trained on ultrasound data* | | | | | | | | | | | | | | | | | | | | | |
| BioMedCLIP [63] | 8.40 | 7.33 | 7.89 | 8.00 | 8.44 | 8.84 | 7.79 | 8.88 | 7.78 | 8.85 | 9.16 | 9.38 | 8.62 | 7.81 | 8.73 | 10.04 | 8.88 | 9.59 | 8.75 | 11.62 | 8.74 |
| LVM-Med [38] | 8.55 | 7.26 | 8.00 | 7.94 | 8.42 | 8.62 | 7.95 | 8.81 | 7.89 | 8.85 | 9.31 | 9.27 | 8.70 | 7.59 | 8.88 | 9.65 | 9.07 | 9.65 | 8.80 | 11.34 | 8.73 |
| US-MoCo [14] | 8.75 | 7.46 | 8.08 | 7.90 | 8.29 | 8.64 | 7.95 | 8.76 | 7.90 | 8.64 | 9.22 | 9.23 | 8.55 | 7.40 | 8.78 | 9.76 | 8.88 | 9.70 | 8.74 | 11.55 | 8.71 |
| US-MAE [27] | 8.31 | 7.11 | 7.74 | 7.61 | 8.30 | 8.23 | 7.69 | 8.55 | 7.70 | 8.56 | **9.02** | 9.05 | 8.35 | 7.26 | 8.47 | 9.67 | 8.55 | 9.46 | 8.45 | 11.05 | 8.46 |
| USFM [31] | 8.34 | 7.14 | 7.74 | 7.69 | 8.28 | 8.51 | 7.62 | 8.38 | 7.69 | 8.41 | **9.02** | 9.00 | 8.20 | 7.18 | 8.33 | 9.38 | 8.51 | 9.37 | 8.39 | 11.14 | 8.42 |
| EchoCLIP [15] | 8.29 | 6.86 | 7.78 | 7.52 | 8.45 | 8.51 | 7.53 | 8.44 | 7.60 | 8.47 | 9.17 | 9.18 | 8.16 | 7.03 | 8.22 | **9.31** | 8.58 | **8.88** | 8.35 | **10.97** | 8.37 |
| EchoWorld | **7.95** | **6.83** | **7.45** | **7.37** | **7.86** | **7.99** | **7.37** | **7.97** | **7.29** | **7.85** | 9.04 | **8.57** | **7.99** | **7.00** | **8.09** | 9.42 | **8.27** | 9.59 | **8.22** | 10.99 | **8.15** |
| *Sequential protocol, using the same visual backbone* | | | | | | | | | | | | | | | | | | | | | |
| US-GuideNet [19] | 7.14 | 6.13 | 7.61 | 7.35 | 8.42 | 7.91 | 6.56 | 7.62 | 7.18 | 7.77 | 8.34 | 8.31 | 7.66 | 6.31 | 7.55 | 7.14 | 8.32 | 8.16 | 8.10 | 10.87 | 7.72 |
| Decision-T [12] | 6.72 | 5.78 | 7.11 | 6.98 | 8.12 | **7.46** | 6.46 | 7.47 | 6.53 | **7.25** | 8.77 | 8.09 | 7.07 | 6.02 | 6.79 | 6.87 | 7.83 | 8.48 | 7.78 | 11.22 | 7.44 |
| Seq-aware [30] | 6.99 | 5.87 | 7.42 | 7.06 | 8.38 | 7.89 | 6.11 | 7.74 | 6.68 | 7.61 | **7.63** | 8.19 | 7.19 | 5.89 | 6.98 | 6.88 | 7.74 | 7.80 | 7.70 | 10.59 | 7.42 |
| EchoWorld | **6.28** | **5.40** | **6.92** | **6.94** | **7.95** | 7.47 | **5.78** | **7.21** | **6.28** | 7.38 | 7.71 | **7.71** | **6.74** | **5.58** | **6.46** | **6.64** | **7.36** | **7.50** | **7.23** | **10.47** | **7.05** |

power of the visual backbones, ensuring that all models are compared using the same prediction head and receive identical supervision signals. We report the mean absolute prediction error averaged across all frames in the test set.

*Sequential protocol*: This protocol simulates an online deployment scenario, where the model predicts probe movements toward *unvisited* planes based on historical visual-motion data[3]. It provides a more comprehensive evaluation of the probe guidance framework's performance. Specifically, for a scan at timestep $t$, the model uses $N$ historical visual-motion pairs $\{\boldsymbol{I}_{t_i}, \boldsymbol{p}_{t_i}\}_{i=1}^{N}$, where $t_i \leq t, t_N = t$. In this setup, the model predicts movements toward planes that have not been visited by timestep $t$ . To ensure a symmetric evaluation, we evaluate the model in both the forward and reversed directions for each scan. The mean absolute prediction error is reported, averaged over every timestep across 30 test scans where all planes are visited. To ensure a fair comparison, the *same* visual backbone (pre-trained using our proposed method) is used for all baselines and our method.

**Baselines.** In the single-frame protocol, we compare EchoWorld against a diverse set of pre-trained models, including general domain vision models such as DeiT [56] and DINOv2 [44], as well as medical vision models like BioMedCLIP [63] and LVM-Med [38], ultrasound-specific models USFM [31] and EchoCLIP [15], as well as two representative self-supervised learning methods applied to

[3]To manage the long frame sequence, we sample $N$ timesteps from the history following a decayed density rate. This allows the model to retain long-term context without losing focus on the current timestep.

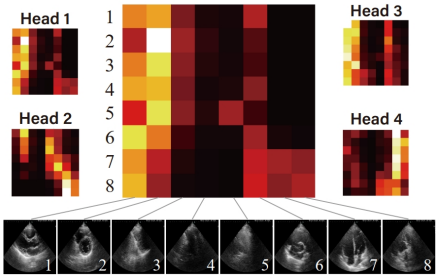


Figure 8. **Visualization of attention scores.**

Table 2. **Effectiveness of the world modeling tasks.** This evaluation uses the single-frame protocol. "Trans." and "Rot." refer to the translation and rotation error averaged over the ten standard planes.

| Spatial | Motion | Trans. | Rot. | Avg |
|---|---|---|---|---|
| ✗ | ✗ | 8.77 | 9.37 | 9.07 |
| ✗ | ✓ | 8.39 | 8.84 | 8.62 |
| ✓ | ✗ | 8.16 | 8.60 | 8.38 |
| ✓ | ✓ | **7.95** | **8.36** | **8.15** |

Table 3. **Ablation study on motion-awareness.** This analysis uses the sequential protocol. "Motion" indicates whether pose information is available.

| Backbone | Motion | Trans. | Rot. | Avg |
|---|---|---|---|---|
| DeiT [56] | ✗ | 8.48 | 8.76 | 8.62 |
| | ✓ | 7.42 | 7.65 | 7.53 |
| EchoWorld | ✗ | 7.91 | 8.06 | 7.98 |
| | ✓ | **6.87** | **7.23** | **7.05** |

our ultrasound dataset, namely US-MoCo and US-MAE, derived from MoCo [14] and MAE [27]. In the sequential protocol, we evaluate EchoWorld against existing probe guidance frameworks, including US-GuideNet [19] and Sequence-aware [30], as well as a sequential decision-making model, Decision Transformer [12]. These baseline frameworks are reproduced using our own dataset to ensure a fair comparison. Additional details on the implementation of these baselines are provided in the appendix.

**Results on the single-frame protocol** are shown in the upper section of Table 1. Our method consistently outperforms all baselines in predicting the probe movement for acquiring the ten standard planes from individual images. EchoWorld achieves the lowest error in 16 out of 20 translation/rotation errors across each subtask, outperforming the best alternative, EchoCLIP, by 0.22 in the averaged mean absolute error. Ultrasound-specific pre-trained models, such as USFM, EchoCLIP, and US-MAE, generally outperform general-purpose models, emphasizing the importance of in-domain transfer.

**Results on sequential protocol** are presented in the lower section of Table 1. EchoWorld surpasses existing probe guidance frameworks while using the same visual backbone initialization across all models. This suggests that the novel guidance prediction module, which incorporates motion-aware attention, enables more effective use of visual-motion data than previous paradigms.

### 5.3. Analytical Results

**Visualization of attention patterns.** To better understand the motion-aware attention mechanism, we visualize the attention scores between a set of 8 ultrasound frames along with their associated motion information. As shown in Figure 8, different attention heads capture diverse attention patterns, where the model tends to focus on high-quality frames (e.g., frames 1 and 2) while ignoring less informative ones (e.g., frames 3, 4, and 5). This demonstrates the model's ability to selectively aggregate useful information for improved probe guidance.

**Visualization of the plane features.** We visualize the image features of the ten standard planes in the dataset using t-SNE. As shown in Figure 9, images representing the same type of plane cluster together, with similar planes appearing closer to each other, indicating that the model learns meaningful semantics relevant to the downstream task.

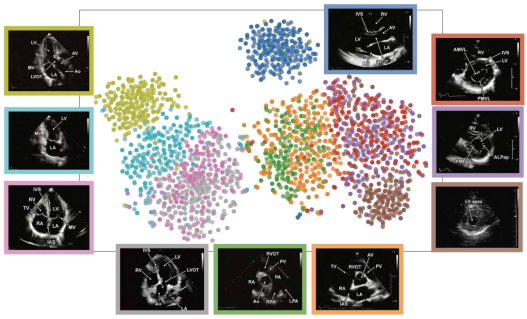

Figure 9. **t-SNE visualization.**

**Ablation study of the world modeling tasks** is shown in Table 2. EchoWorld benefits from combining both spatial and motion modeling, leading to a significant improvement in representation quality.

**Ablation study on motion-awareness** is shown in Table 3. Our results show that leveraging motion signals significantly reduces guidance errors. Additionally, when using the pre-trained backbone, the motion-aware guidance module provides a further performance boost, demonstrating the coherence and effectiveness of the proposed framework.

## 6. Conclusion

This paper proposed EchoWorld, a motion-aware world modeling framework designed for probe guidance in echocardiography. By integrating spatial and motion-aware world modeling tasks, our pre-trained encoder captures rich representations of ultrasound knowledge, encompassing anatomical structures as well as motion-driven dynamic visual changes. A motion-aware attention mechanism further enhances probe guidance by seamlessly incorporating the visual-motion features. Our model outperforms existing pre-trained models and frameworks in predicting probe movements for acquiring standard planes. Our work provides a principled approach to modeling medical images, which considers both the medical knowledge and the inherent dynamics of imaging processes, advancing machine learning models for embodied medical applications.

## Acknowledgements

The work is supported in part by the National Key R&D Program of China under Grant 2024YFB4708200.

## References

[1] Decart AI. Oasis: A universe in a transformer. https://www.decart.ai/articles/oasis-interactive-ai-video-game-model. Accessed: Nov, 2024. 3

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *arXiv preprint arXiv:2405.12399*, 2024. 3

[3] Abdoul Aziz Amadou, Vivek Singh, Florin C Ghesu, Young-Ho Kim, Laura Stanciulescu, Harshitha P Sai, Puneet Sharma, Alistair Young, Ronak Rajani, and Kawal Rhode. Goal-conditioned reinforcement learning for ultrasound navigation guidance. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 319–329. Springer, 2024. 3

[4] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 3, 4

[5] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, pages 1298–1312. PMLR, 2022. 4

[6] Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR, 2023.

[7] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023. 3

[8] Florian Bordes, Randall Balestriero, and Pascal Vincent. High fidelity visualization of what your self-supervised representation knows about. *arXiv preprint arXiv:2112.09164*, 2021. 6

[9] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023. 3

[10] Arthur Earl Bryson. *Applied optimal control: optimization, estimation and control*. Routledge, 2018. 2

[11] Eduardo F Camacho, Carlos Bordons, Eduardo F Camacho, and Carlos Bordons. *Constrained model predictive control*. Springer, 2007. 2

[12] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021. 7, 8

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 5

[14] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021. 7, 8

[15] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024. 3, 7

[16] Kenneth James Williams Craik. *The nature of explanation*. CUP Archive, 1967. 2, 4

[17] Xiaolong Deng, Huisi Wu, Runhao Zeng, and Jing Qin. Memsam: Taming segment anything model for echocardiography video segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9622–9631, 2024. 3

[18] 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. *arXiv preprint arXiv:2010.11929*, 2020. 7

[19] Richard Droste, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Automatic probe movement guidance for freehand obstetric ultrasound. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 583–592. Springer, 2020. 1, 3, 7, 8

[20] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024. 3

[21] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024. 3

[22] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018. 2

[23] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. 2, 4

[24] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 3

[25] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020. 2

[26] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3

[27] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4, 7, 8

[28] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 3

[29] Haojun Jiang, Zhenguo Sun, Ning Jia, Meng Li, Yu Sun, Shaqi Luo, Shiji Song, and Gao Huang. Cardiac copilot: Automatic probe guidance for echocardiography with world model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 190–199. Springer, 2024. 1, 3

[30] Haojun Jiang, Zhenguo Sun, Yu Sun, Ning Jia, Meng Li, Shaqi Luo, Shiji Song, and Gao Huang. Sequence-aware pre-training for echocardiography probe guidance. *arXiv preprint arXiv:2408.15026*, 2024. 3, 7, 8

[31] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical Image Analysis*, page 103202, 2024. 3, 7

[32] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 3

[33] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[34] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1), 2022. 2, 4

[35] Keyu Li, Ang Li, Yangxin Xu, Huahua Xiong, and Max Q-H Meng. Rl-tee: Autonomous probe guidance for transesophageal echocardiography based on attention-augmented deep reinforcement learning. *IEEE Transactions on Automation Science and Engineering*, 21(2):1526–1538, 2023. 3

[36] Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2024. 3

[37] Qianhui Men, Clare Teng, Lior Drukker, Aris T Papageorghiou, and J Alison Noble. Gaze-probe joint guidance with multi-task learning in obstetric ultrasound scanning. *Medical image analysis*, 90:102981, 2023. 3

[38] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[39] Carol Mitchell, Peter S Rahko, Lori A Blauwet, Barry Canaday, Joshua A Finstuen, Michael C Foster, Kenneth Horton, Kofo O Ogunyankin, Richard A Palma, and Eric J Velazquez. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the american society of echocardiography. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019. 3

[40] Akhil Narang, Richard Bae, Ha Hong, Yngvil Thomas, Samuel Surette, Charles Cadieu, Ali Chaudhry, Randolph P Martin, Patrick M McCarthy, David S Rubenson, et al. Utility of a deep-learning algorithm to guide novices to acquire echocardiograms for limited diagnostic use. *JAMA cardiology*, 6(6):624–632, 2021. 1, 3

[41] Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and rt-x models. *arXiv preprint arXiv:2310.08864*, 2023. 3

[42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[43] OpenAI. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/. Accessed: Feb, 2024. 3

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[45] David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, Euan A Ashley, et al. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. 3

[46] Gregory A Roth, Catherine Johnson, Amanuel Abajobir, Foad Abd-Allah, Semaw Ferede Abera, Gebre Abyu, Muktar Ahmed, Baran Aksut, Tahiya Alam, Khurshid Alam, et al. Global, regional, and national burden of cardiovascular diseases for 10 causes, 1990 to 2015. *Journal of the American college of cardiology*, 70(1):1–25, 2017. 1

[47] RunwayML. Introducing general world models. https://research.runwayml.com/introducing-general-world-models. Accessed: May, 2023. 2

[48] Sigbjorn Sabo, David Pasdeloup, Hakon Neergaard Pettersen, Erik Smistad, Andreas Østvik, Sindre Hellum Olaisen, Stian Bergseng Stølen, Bjørnar Leangen Grenne, Espen Holte, Lasse Lovstakken, et al. Real-time guidance by deep learning of experienced operators to improve the standardization of echocardiographic acquisitions. *European Heart Journal-Imaging Methods and Practice*, 1(2): qyad040, 2023. 1, 3

[49] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018. 6

[50] Yuuki Shida, Souto Kumagai, Ryosuke Tsumura, and Hiroyasu Iwata. Automated image acquisition of parasternal long-axis view with robotic echocardiography. *IEEE Robotics and Automation Letters*, 2023. 1, 3

[51] Yuuki Shida, Masami Sugawara, Ryosuke Tsumura, Haruaki Chiba, Tokuhisa Uejima, and Hiroyasu Iwata. Diagnostic posture control system for seated-style echocardiography robot. *International Journal of Computer Assisted Radiology and Surgery*, 18(5):887–897, 2023. 3

[52] Peige Song, Zhe Fang, Hanyu Wang, Yutong Cai, Kazem Rahimi, Yajie Zhu, F Gerald R Fowkes, Freya JI Fowkes, and Igor Rudan. Global and regional prevalence, burden, and risk factors for carotid atherosclerosis: a systematic review, meta-analysis, and modelling study. *The Lancet Global Health*, 8 (5):e721–e729, 2020. 1

[53] Hari Sowrirajan, Jingbo Yang, Andrew Y Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. In *Medical Imaging with Deep Learning*, pages 728–744. PMLR, 2021. 5

[54] David Stojanovski, Uxio Hermida, Pablo Lamata, Arian Beqiri, and Alberto Gomez. Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pages 34–43. Springer, 2023. 3

[55] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 6

[56] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 7, 8

[57] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 3

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 6

[59] Jian Wang, Xin Yang, Xiaohong Jia, Wufeng Xue, Rusi Chen, Yanlin Chen, Xiliang Zhu, Lian Liu, Yan Cao, Jianqiao Zhou, et al. Thyroid ultrasound diagnosis improvement via multi-view self-supervised learning and two-stage pretraining. *Computers in Biology and Medicine*, 171:108087, 2024. 3

[60] Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, and Gao Huang. Efficienttrain: Exploring generalized curriculum learning for training visual backbones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5852–5864, 2023. 4

[61] Zhongnuo Yan, Xin Yang, Mingyuan Luo, Jiongquan Chen, Rusi Chen, Lian Liu, and Dong Ni. Fine-grained context and multi-modal alignment for freehand 3d ultrasound reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 340–349. Springer, 2024. 3

[62] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14662–14672, 2024. 3

[63] 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. 7