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

Supplementary Material

A. Dataset

The echocardiography dataset used in this study was collected during routine clinical examinations, where certified sonographers performed ultrasound scans (M5S probe, GE Vivid E7 machine) using a probe mounted on a Franka Panda robot arm. During each scan, both the ultrasound videos (30 fps) and the corresponding probe poses were simultaneously recorded. All subjects in the dataset were healthy adult males. The data collection process was conducted in compliance with ethical guidelines and was reviewed and approved by the relevant institutional ethics committee.

This study utilizes a subset of 356 scans curated from the dataset, comprising approximately one million images and corresponding ultrasound probe poses. Each scan lasts several minutes, during which the sonographer maneuvers the probe and examines the heart from various views.

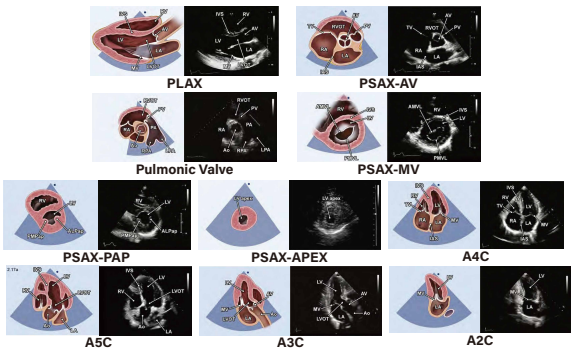

Figure A1. **Ten standard planes.**

For the probe guidance task, we consider ten target standard planes recommended by the American Society of Echocardiography [13], as shown in Figure A1. These planes include: Parasternal Long-Axis (PLAX), Parasternal Short-Axis Aortic Valve (PSAX-AV), Pulmonic Valve (PSAX-PV), Mitral Valve (PSAX-MV), Papillary Muscles (PSAX-PAP), Level of Apex (PSAX-APEX), Apical Four-Chamber (A4C), Apical Five-Chamber (A5C), Apical Three-Chamber (A3C), and Apical Two-Chamber (A2C). Professionals manually annotate the timestamps and frames corresponding to these planes, which serve as the ground truth for the probe guidance task. The dataset is divided into separate training (284 scans) and testing (72 scans) sets, with no overlap of individuals between the two.

B. Tasks and Baselines

The probe guidance task in our study involves predicting the probe's movement toward ten standard planes. The prediction can rely on either a single image or incorporate past visual-motion data. Specifically, in an ultrasound scan comprising T frames, represented as $\{\boldsymbol{I}_t, \boldsymbol{p}_t\}_{t=1}^{T}$, experts identify the timestamps at which the ten standard planes are observed, denoted as $s_1, s_2, \ldots, s_{10}$. For each timestep t with image $\boldsymbol{I}_t$ and corresponding pose $\boldsymbol{p}_t$, the relative pose to the k-th standard plane is computed as $\boldsymbol{a}_t^{(k)} = \boldsymbol{p}_{s_k} \cdot \boldsymbol{p}_t^{-1}$. The model's objective is to predict these movements $\boldsymbol{a}_t$ based on the available visual-motion data.

The probe pose is represented in six degrees of freedom $\boldsymbol{p} \in \mathbb{R}^6$, where the first three components represent translations (x,y,z) in millimeters, and the last three correspond to rotations (yaw, pitch, roll) in degrees. For model evaluation, we calculate the mean absolute error separately for translation and rotation components, as detailed in Table 1.

We employ two evaluation protocols (single-frame and sequential) in our study, as described in Section 5.2. Below, we provide a detailed introduction to each protocol.

B.1. Single-Frame Protocol

In the single-frame protocol, the model predicts the probe's movement toward all ten standard planes using a single ultrasound image as input. This setup evaluates the representation quality of pre-trained visual models in a cost-efficient manner. Two-layer MLPs are appended to the pre-trained backbones, and the entire model undergoes full fine-tuning. The evaluation metric is computed as the average error across all frames in the test set. To improve evaluation efficiency, the frame rate is reduced to 6 fps.

In this protocol, we evaluate the performance of our visual encoder, pre-trained using world modeling tasks, against a diverse selection of pre-trained models. These include DeiT [15], DINOv2 [14], BioMedCLIP [16], LVM-Med [12], US-MoCo [4], US-MAE [9], USFM [11], EchoCLIP [5]. For consistency, we use the ViT-Small variant of each method whenever available. Below, we provide an overview of these baselines:

- **DeiT** [15] is a family of vision transformers trained on the ImageNet dataset [6].
- **DINOv2** [14] is a state-of-the-art self-supervised vision foundation model trained on a wide range of general-domain images. The training algorithm mainly follows DINO [2] and iBOT [17].

- **BioMedCLIP** [16] is a multimodal biomedical foundation model pre-trained on 15 million medical image-text pairs using contrastive learning. We utilize only the visual encoder component of this model for our comparisons.
- **LVM-Med** [12] employs a graph-matching formulation for contrastive learning, enabling it to integrate multiple medical imaging modalities, including ultrasound, into a single versatile framework.
- **US-MoCo** [4] is an adaptation of the MoCo framework to our dataset. MoCo employs a momentum encoder to create a dynamic dictionary for stable and effective representation learning. We pre-train a ViT-Small model on our ultrasound dataset using the MoCov3 codebase, training for 150 epochs with a learning rate of 1.5×10^{-4}, weight decay of 0.1, and batch size of 1024.
- **US-MAE** [9] is an adaptation of the MAE framework to our dataset. MAE is an encoder-decoder framework for mask image modeling. For this adaptation, we train a ViT-Small with a four-layer decoder, a masking ratio of 0.75, over 300 epochs. The training setup includes a learning rate of 6×10^{-4}, weight decay of 0.05, and batch size of 1024.
- **USFM** [11] is an ultrasound-specific vision foundation model trained on over 2 million ultrasound images using a spatial-frequency dual mask modeling approach.
- **EchoCLIP** [5] is a multimodal foundation model for echocardiogram interpretation. The model is trained on more than 1 million ultrasound image-text pairs using contrastive learning. We utilize only the visual encoder component of this model for our comparisons.

B.2. Sequential Protocol

The sequential protocol simulates a deployment scenario, where the model predicts the probe's movement toward unvisited planes based on past visual-motion data up to the current timestep (visited planes are excluded from the prediction error calculation). It provides a more holistic assessment of probe guidance frameworks. In this setting, we use our pre-trained visual encoder as the backbone for all baselines. Specifically, at the timestep t of a scan, the model uses historical visual-motion data before t to predict the standard planes that are yet to be visited. The history data $\mathcal{H}_t$ and the target plane indices $\mathcal{K}_t$ are defined by:

$$\begin{aligned} \mathcal{H}_t &= \{(\boldsymbol{I}_{t'}, \boldsymbol{p}_{t'}) | t' < t\}, \\ \mathcal{K}_t &= \{k | s_k \geq t\}, \end{aligned} \tag{1}$$

where s_k is the timestep when the k-th plane is visited. To construct the model inputs, we sample N visual-motion pairs $\{\boldsymbol{I}_{t_i}, \boldsymbol{p}_{t_i}\}_{i=1}^{N}$ from $\mathcal{H}_t$ using a decayed density sampling rate. This approach ensures that recent observations are prioritized while retaining a representative selection of past data. The sampled timesteps t_i are computed as:

$$t_i = \text{Round}\left(t + \frac{t}{\alpha N} \log \frac{i}{N}\right), \quad i = 1, \ldots, N, \tag{2}$$

where α is a scaling factor. By default, we sample $N = 8$ frames from the history with $\alpha = 0.4$. If the history contains fewer than eight frames, we allow repeated sampling to meet the required count. To ensure a symmetric evaluation, we assess the model in both forward and reverse directions for each scan. In the reverse direction, the scan begins at the last frame. The historical data $\widetilde{\mathcal{H}}_t$ and target plane indices $\widetilde{\mathcal{K}}_t$ at timestep t are defined as:

$$\begin{aligned} \widetilde{\mathcal{H}}_t &= \{(\boldsymbol{I}_{t'}, \boldsymbol{p}_{t'}) | t' \geq t\}, \\ \widetilde{\mathcal{K}}_t &= \{k | s_k < t\}. \end{aligned} \tag{3}$$

The final error metric is averaged over both forward and reverse directions and all timesteps across the scans. For computational efficiency, the frame rate is reduced to 3 fps during this evaluation.

In this protocol, we evaluate the complete EchoWorld framework, which incorporates the proposed motion-aware attention mechanism, by comparing it against existing probe guidance frameworks. These include US-GuideNet [7], Decision-Transformer [3], and Sequence-aware Pre-training [10]. To ensure a fair comparison and isolate the impact of our motion-aware modeling, all baselines use the same visual encoder. The visual encoder extracts average-pooled image features, which are subsequently passed to the respective probe guidance frameworks. Below, we provide detailed descriptions of these baselines:

- **US-GuideNet** [7] is originally designed for freehand obstetric ultrasound probe guidance. In our implementation, we adopt its model design, which processes sequential inputs in the form:

$$\{\boldsymbol{I}_1, \boldsymbol{p}_{1\to2}, \boldsymbol{I}_2, \boldsymbol{p}_{2\to3}, \boldsymbol{I}_3, \ldots, \boldsymbol{I}_N\}. \tag{4}$$

  Here, $\boldsymbol{p}_{i\to i+1}$ denotes probe movements between consecutive frames. Visual and motion features are projected and concatenated before being aggregated using a gated recurrent unit (GRU).
- **Decision-Transformer** [3] models trajectories within a Markov Decision Process using a causal transformer. For our task, we adapt this architecture by feeding interleaved states (images) and actions (probe movements) using the same input structure as Equation (4). The interleaved sequence is passed through a two-layer causal transformer, with the output of the final token feeding into a guidance prediction head for downstream tasks.
- **Sequence-aware Pre-training** [10] utilizes a bidirectional transformer to process interleaved visual-motion sequences, adhering to the same input format as Equation

Algorithm 1 PyTorch-style pseudocode for motion-aware attention.

```
# B: batch size
# N: number of frames
# D_img: dimensionality of the image features
# D_mo: dimensionality of the motion features
# D: dimensionality of attention features
# x_img:  image features  shaped (B, N, D_img)
# x_motion:  motion features shaped (B, N, N, D_mo)

def motion_aware_attn(x_img, x_motion):
    # expand image features
    x_img_exp = x_img.unsqueeze(1).expand(B,N,N,D_img)

    # compute query, key, and value
    Q = mlp_q(x_img) # BxNxD
    K = mlp_k(concat(x_img_exp, x_motion)) # BxNxNxD
    V = mlp_v(concat(x_img_exp, x_motion)) # BxNxNxD

    # perform attention
    logits = einsum('bid,bijd->bij', Q,K) / (D ** 0.5)
    attn = softmax(logits, dim=-1)
    return einsum('bij,bijd->bid', attn, V)
```

(4). The model is pre-trained using a visual-motion mask modeling strategy to enhance historical data aggregation. During fine-tuning, an extra mask token is appended to the sequence for probe movement prediction.

C. Implementation Details

C.1. Pre-training

Architecture and optimization. EchoWorld is pre-trained from scratch by jointly performing spatial and motion world modeling. The context encoder is a ViT-S/16, while the target encoder is an exponential moving average (EMA) of the context encoder with a starting decay rate of 0.996, which gradually increases to 1.0 following a cosine schedule. The predictor is a 6-layer transformer with a width of 384. Input images are resized to 224×224. The model is optimized using the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, an initial learning rate of 10^{-3}, and a weight decay of 0.05. Training spans 300 epochs, with a 40-epoch linear warm-up followed by cosine decay. The default batch size is 1024, and training takes approximately 14 hours on four A100 GPUs.

Spatial world modeling. Following [1], the context image is masked using four rectangular blocks with scales ranging from $(0.15, 0.2)$. The visible regions are further reduced by up to 15%, increasing the task's difficulty. Only visible patches are processed by the context encoder, whereas the target encoder takes the entire image as input. In the predictor, mask tokens, enriched with positional encodings corresponding to the masked patches, are concatenated with context tokens. A smoothed L1 loss is computed between the predicted and target outputs at the masked locations.

Motion world modeling. We randomly sample two frames $\boldsymbol{I}_a, \boldsymbol{I}_b$ along with their respective poses $\boldsymbol{p}_a, \boldsymbol{p}_b$ from a scan and compute their relative pose difference $\boldsymbol{p}_{a\to b} = \boldsymbol{p}_b \cdot \boldsymbol{p}_a^{-1}$. Frame $\boldsymbol{I}_a$ is used as input to the context encoder, while frame $\boldsymbol{I}_b$ serves as the target. The motion encoder A_ψ is a two-layer MLP with a hidden dimension of 384, producing motion feature $\boldsymbol{z}_{a\to b} = A_\psi(\boldsymbol{p}_{a\to b})$. These features are embedded into a mask token and concatenated with context tokens before being passed to the predictor. The predictor generates $\hat{h}_y$, a prediction of the average-pooled target feature h_y. Before computing the InfoNCE loss, $\hat{h}_y$ and h_y are projected using projectors P and P', where P' is an EMA of P. For simplicity, we skip the projector in Equation (4) of the main paper. The loss, including the projector, is defined as:

$$\mathcal{L}_{motion} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(P(\hat{h}_{y_i})^\top \cdot P'(h_{y_i})/\tau)}{\sum_j \exp(P(\hat{h}_{y_i})^\top \cdot P'(h_{y_j})/\tau)}, \tag{5}$$

where B is the batch size and τ is the temperature (set to 0.1 by default). The loss can be symmetrized by swapping the context and target roles.

Joint modeling. The integration of spatial and motion world modeling follows a unified pipeline. Specifically, for the frames $\boldsymbol{I}_a$ and $\boldsymbol{I}_b$ used in motion modeling, some regions in the context frame $\boldsymbol{I}_a$ are masked. The predictor simultaneously performs two tasks: (1) reconstructing masked regions in the context frame and (2) predicting features of the target frame based on motion information. The predictions and targets for these tasks are defined as:

$$\begin{aligned} h_x &= f_\theta(\mathrm{Mask}(\boldsymbol{I}_a, M)), \\ \hat{h}_y^{\text{spatial}} &= g_\phi(h_x + p_x; \{m + \mathrm{PE}(c)\}_{c\in M}), \\ h_y^{\text{spatial}} &= \{f'_{\theta'}(\boldsymbol{I}_a)_c\}_{c\in M}, \\ \hat{h}_y^{\text{motion}} &= g_\phi(h_x; m + \boldsymbol{z}_{a\to b}), \\ h_y^{\text{motion}} &= \mathrm{AvgPool}(f'_{\theta'}(\boldsymbol{I}_b)), \end{aligned} \tag{6}$$

where $\hat{h}_y^{\text{spatial}}, h_y^{\text{spatial}}$ are prediction and target for spatial modeling, and $\hat{h}_y^{\text{motion}}, h_y^{\text{motion}}$ are for motion modeling. The total loss combines both objectives: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spatial}} + \lambda\mathcal{L}_{\text{motion}}$, where $\lambda = 0.1$ balances the scale of the two losses.

C.2. Fine-tuning

Motion-aware attention. Algorithm 1 provides the pseudocode for the proposed motion-aware attention mechanism. The pre-trained visual encoder f_θ and motion encoder A_ψ extract visual and motion features, $\boldsymbol{h}_i$ and $\boldsymbol{z}_{i\to j}$, for frames $i, j \in [1, N]$. Two MLPs process their concatenation to generate keys $K_j^{(i)}$ and values $V_j^{(i)}$ as follows:

$$K_j^{(i)} = \mathrm{MLP}_k(\boldsymbol{h}_j, \boldsymbol{z}_{i\to j}),\ V_j^{(i)} = \mathrm{MLP}_v(\boldsymbol{h}_j, \boldsymbol{z}_{i\to j}). \tag{7}$$

Queries are derived from the image features using another MLP: $Q_i = \text{MLP}_q(\boldsymbol{h}_i)$. The model applies scaled dot-product attention with four attention heads and a hidden dimension of 384. The resulting attention outputs are passed through ten independent MLPs to predict probe movements to ten standard planes relative to the current pose.

Optimization. The model is optimized using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of 1×10^{-4}. Training uses 15,000 iterations with a batch size of 256 for single-frame and 64 for sequential protocols. Additional settings include weight decay of 0.05, drop path of 0.1, layer-wise learning rate decay of 0.65, and random brightness/contrast augmentations.

C.3. Visualizations

World model predictor outputs (Figure 7). To better understand the predictor outputs of our world model, we train a diffusion model to reconstruct target pixel values conditioned on the representation $\hat{h}_y$ produced by the predictor. This guidance representation is first projected to a 512-dimensional vector, which is then integrated into the diffusion model via conditional batch normalization layers [8]. For spatial world modeling, the diffusion model is conditioned on the average-pooled predictor outputs corresponding to the masked regions. For motion world modeling, the diffusion model takes the predictor output vector as its conditioning signal. We train separate diffusion models for the two world modeling tasks, with both models trained for 300,000 iterations and generating images at resolution 128×128.

Analysis of attention scores (Figure 8). We evaluate the proposed motion-aware attention mechanism by visualizing attention scores across a set of eight visual-motion pairs. Some of these pairs include noisy frames with minimal usable information. For this input, we visualize the 8×8 attention score matrices across all four attention heads. Each matrix entry, located at the i-th row and j-th column, represents the attention score of query i attending to key j.

Analysis of plane features (Figure 9). We extract average-pooled representations of all standard plane images identified by professionals in the training set. These representations are visualized in a 2D space using t-SNE. The visualization highlights how well the model clusters images of similar planes.

References

[1] 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

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[3] 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. 2

[4] 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. 1, 2

[5] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, pages 1–8, 2024. 1, 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[7] 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. 2

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *arXiv preprint arXiv:1610.07629*, 2016. 4

[9] 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. 1, 2

[10] 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. 2

[11] 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. 1, 2

[12] 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. 1, 2

[13] 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 echocardiogra-

phy. *Journal of the American Society of Echocardiography*, 32(1):1–64, 2019. 1

[14] 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. 1

[15] 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. 1

[16] 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. 1, 2

[17] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *arXiv preprint arXiv:2111.07832*, 2021. 1