# VA-Adapter: Adapting Ultrasound Foundation Model to Echocardiography Probe Guidance

Teng Wang[1*], Haojun Jiang[1*‡], Yuxuan Wang[2], Zhenguo Sun[3], Yujiao Deng[4], Shiji Song[1], and Gao Huang[1†]

[1] Department of Automation, BNRist, Tsinghua University, Beijing, China
[2] School of Computer Science and Technology, Xidian University, Xi'an, China
[3] Beijing Academy of Artificial Intelligence, Beijing, China
[4] Chinese PLA General Hospital, Beijing, China
{t-wang25, jhj20}@mails.tsinghua.edu.cn, gaohuang@tsinghua.edu.cn

**Abstract.** Echocardiography is a critical tool for detecting heart diseases, yet its steep operational difficulty causes a shortage of skilled personnel. Probe guidance systems, which assist in acquiring high-quality images, offer a promising solution to lower this operational barrier. However, robust probe guidance remains challenging due to significant individual variability. This variability manifests as differences in low-level features within two-dimensional (2D) images, which complicates image feature understanding, and differences in individual three-dimensional (3D) structures, which poses challenges for precise navigation. To address these challenges, we first propose leveraging the robust image representations learned by ultrasound foundation models from vast datasets. Yet, applying these models to probe navigation is non-trivial due to their lack of understanding of individual 3D structures. To this end, we meticulously design a Vision-Action Adapter (VA-Adapter) to online inject the capability of understanding individual 3D structures. Specifically, by embedding the VA-Adapter into the foundation model's image encoder, the model can infer cardiac anatomy from historical vision-action sequences, mimicking the cognitive process of a sonographer. Extensive experiments on a dataset with over 1.31M samples demonstrate that the VA-Adapter outperforms strong probe guidance models while requiring approximately 33 times fewer trained parameters. Code is available at https://github.com/LeapLabTHU/VA-Adapter.

**Keywords:** Echocardiography · Adapter · Probe Guidance · Foundation Model · Sequence Modeling.

## 1 Introduction

Cardiovascular diseases have become a significant factor affecting human lifespan [22]. Echocardiography is a commonly used imaging technique for diagnosing cardiovascular diseases, allowing the observation of the health conditions of

* Equal contribution. ‡ Guided this work. † Corresponding author.

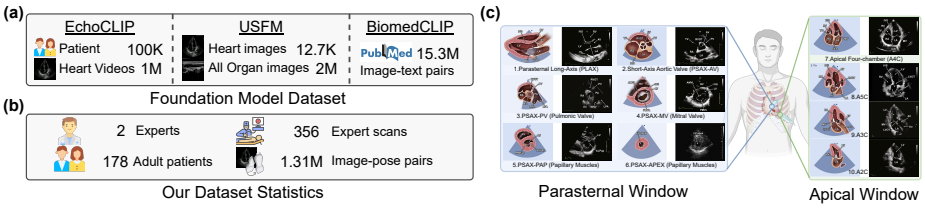


Fig. 1: Illustration of the dataset. (a) Large-scale diagnostic foundation model dataset. (b) Our dataset statistic. (c) Standard planes (images from [18]).

heart chambers, valves, and blood vessels [18]. With technological advancements, AI-driven echocardiography diagnostic models [5, 7, 21] have demonstrated remarkable capabilities. As shown in Fig. 1(a), representative models include EchoCLIP [5], pre-trained on over one million paired cardiac ultrasound videos and expert reports; USFM [14], pre-trained on over two million ultrasound images across 12 organs; and BiomedCLIP [28], pre-trained on 15.3 million image–text pairs from 4.4 million PubMed articles. All three models demonstrate strong capabilities in interpreting cardiac ultrasound images.

Undoubtedly, the prerequisite for these powerful diagnostic models to function effectively is the availability of high-quality ultrasound images. However, due to the inherently high operational difficulty of cardiac ultrasound, it takes years of training for a beginner to master the technique, resulting in a scarcity of skilled professionals. Therefore, leveraging AI technology to assist in cardiac ultrasound scanning is a crucial research direction.

In recent years, researchers [2,6,10–13,15,19,25,27] have developed AI-driven probe guidance systems to acquire higher-quality ultrasound images. Droste et al. [6] proposed US-GuideNet for fetal plane scanning. Narang et al. [19] collected large-scale echocardiography scanning data and trained a CNN-based guidance system from scratch, but it is commercial and closed-source. Li et al. [15] used cardiac CT as a simulation environment and applied reinforcement learning to learn guidance policies. Despite progress, these works study probe guidance separately from diagnosis and do not leverage advances in diagnostic foundation models [5, 14, 28] (e.g., EchoCLIP). Since both scanning and diagnosis require understanding cardiac ultrasound structures and making decisions, we hypothesize that diagnostic-model advances can also improve probe guidance.

In this paper, we aim to build upon the ultrasound foundation model's ability to interpret cardiac ultrasound images by equipping it with the capability to understand three-dimensional (3D) cardiac structures and reason about probe adjustment actions. First, to preserve the basic capabilities learned by foundation model from large-scale data as much as possible, we employ a parameter-efficient fine-tuning strategy called Adapter, which freezes the foundation model's image encoder and only optimizes the parameters within the adapter. Further, to better exploit individual 3D anatomy, we propose a Vision–Action Adapter (VA-Adapter) that encodes vision–action sequences and learns individual-specific car-

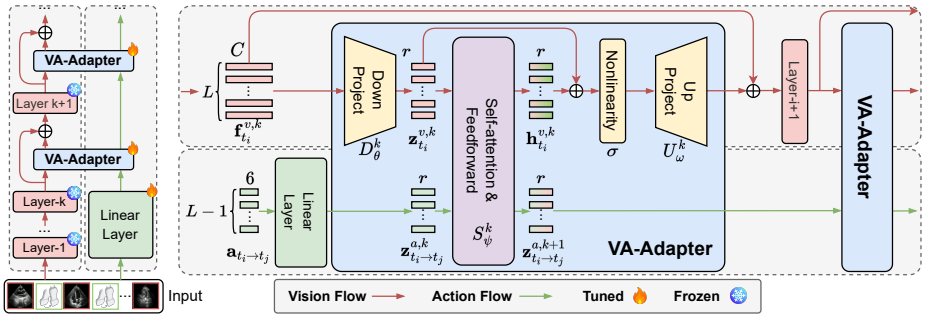


Fig. 2: Illustration of the architecture of the VA-Adapter. The left side shows that we insert VA-Adapter into the deep layers of foundation models, and the right side shows the internal structure of VA-Adapter.

diac structure from them. Specifically, VA-Adapters are inserted into deeper encoder layers of foundation model, where features are more task-relevant [26]. Finally, extensive experiments show that the VA-Adapter equips diagnostic foundation models with better probe guidance capability at a low training cost.

## 2 Method

In this section, We first describe the cardiac ultrasound scanning dataset, then introduce our Vision–Action Adapter.

### 2.1 Dataset Acquisition

The dataset used in our work was collected from 178 adult subjects and includes 356 expert scanning trajectories, totaling 1.31 million image-action pairs. The data was gathered by two senior sonographers with over 10 years of experience. They performed continuous scans of 10 standard echocardiographic planes (Fig. 1(c)) using an ultrasound probe attached to the end of a robotic arm. The data collection system recorded real-time images acquired by the probe and the corresponding probe pose data. This created a scanning sequence $\{(\mathbf{I}_t, \mathbf{p}_t)\}_{t=1}^{T}$, where $\mathbf{I}_t \in \mathbb{R}^{3\times H\times W}$ is the ultrasound image at time $t$, and $\mathbf{p}_t \in \mathbb{SE}(3)$ is the corresponding probe's 6D pose (3D position $\mathbf{P} \in \mathbb{R}^3$ and 3D orientation $\mathbf{R} \in \mathbb{SO}(3)$). During the scan, the sonographer marked the standard views. Then we can calculate the relative motion of any image $\mathbf{I}_i$ to the standard view $\mathbf{I}_j$ : $\mathbf{a}_{i\to j} = \mathbf{T}_{p_i}^{-1}\mathbf{T}_{p_j}$, where $\mathbf{T}_{p_i}$ and $\mathbf{T}_{p_j}$ represent the transformation matrices corresponding to the probe poses $\mathbf{p}_i$ and $\mathbf{p}_j$. Then we use this motion value as the supervisory signal for the ultrasound probe guidance task. Notably, sonographers sometimes pause at a probe position to observe cardiac dynamics. As a result, the dataset includes many images from the same probe position but different cardiac phases. These frames share the same motion label, providing implicit supervision that encourages robustness to cardiac-cycle variation.

### 2.2 Vision-Action Adapter

Echocardiography probe guidance is an emerging area, but progress is limited by data-collection challenges. In contrast, ultrasound image understanding and diagnosis have advanced rapidly, with recent foundation models achieving strong baseline performance. Since probe guidance also requires understanding ultrasound structures, we propose the Vision-Action Adapter to leverage these foundation models for probe guidance. Next, we detail our approach in Fig. 2.

**Input.** Given a frame $\mathbf{I}_t \in \mathbb{R}^{3\times H\times W}$ from a scan, we construct an input sequence of length $L$ via segmental sampling [24]. Specifically, we split the preceding trajectory into $L-1$ equal temporal segments and randomly sample one frame (and its pose) from each segment. The sampled frames are sorted by timestamp and concatenated with the current frame $\mathbf{I}_t$ as the last element $\mathbf{I}_{t_L}$. Compared with using $L$ consecutive frames, segmental sampling yields larger inter-frame motion and more diverse anatomical viewpoints, which helps the model capture richer 3D structural cues. For reference, segmental sampling yields an average absolute inter-frame motion of [21.8, 18.7, 14.0] mm and [12.6, 23.3, 40.9]° between adjacent frames, indicating substantial variation in both position and orientation. We then compute relative actions $\mathbf{a}_{t_i\to t_j} \in \mathbb{R}^6$ from the corresponding poses, forming the input trajectory: $[\mathbf{I}_{t_1}, \mathbf{a}_{t_1\to t_2}, \cdots, \mathbf{I}_{t_{L-1}}, \mathbf{a}_{t_{L-1}\to t_L}, \mathbf{I}_{t_L}]$.

**Forward Propagation.** We insert VA-Adapters into the latter part of the foundation model's vision encoder, where features are more task-relevant [26]. For CNN encoders (e.g., EchoCLIP), VA-Adapters are placed between late-stage blocks; for Transformer encoders (e.g., USFM, BiomedCLIP), two VA-Adapters are inserted in each selected block (after attention and after MLP). During training, only VA-Adapters are updated while the backbone remains frozen to preserve its learned knowledge. Given an image $\mathbf{I}_{t_i}$, after the first $k$ vision layers we obtain $\mathbf{f}_{t_i}^{v,k} \in \mathbb{R}^C$ (for CNNs, global average pooling is applied). We project visual features and actions to the bottleneck space and add timestep embeddings:

$$\mathbf{z}_{t_i}^{v,k} = D_\theta^k(\mathbf{f}_{t_i}^{v,k}) + \mathbf{T}_i, \quad \mathbf{z}_{t_i\to t_j}^{a,k} = A_\phi^k(\mathbf{a}_{t_i\to t_j}) + \mathbf{T}_i, \quad \mathbf{z}_{t_i}^{v,k}, \mathbf{z}_{t_i\to t_j}^{a,k} \in \mathbb{R}^r, \tag{1}$$

where $\mathbf{T} \in \mathbb{R}^{L\times r}$ is timestep embedding, and $r$ is the bottleneck dimension. The interleaved vision–action tokens are then processed by a vision–action interaction module $S_\psi^k$, designed to extract underlying cardiac structure information:

$$[\mathbf{h}_{t_1}^{v,k}, \mathbf{z}_{t_1\to t_2}^{a,k+1}, \cdots, \mathbf{h}_{t_L}^{v,k}] = S_\psi^k([\mathbf{z}_{t_1}^{v,k}, \mathbf{z}_{t_1\to t_2}^{a,k}, \cdots, \mathbf{z}_{t_{L-1}}^{v,k}, \mathbf{z}_{t_{L-1}\to t_L}^{a,k}, \mathbf{z}_{t_L}^{v,k}]). \tag{2}$$

The action tokens are forwarded to the next adapter layer, while visual tokens are mapped back to the backbone feature space with a residual connection:

$$\mathbf{z}_{t_i}^{v,k+1} = U_\omega^k\Big(\sigma(\mathbf{h}_{t_i}^{v,k} + \mathbf{z}_{t_i}^{v,k})\Big) + \mathbf{f}_{t_i}^{v,k}. \tag{3}$$

### 2.3 Task Prediction Head

After passing through all layers (assuming the network has $M$ layers in total), we obtain the final interleaved sequence features:

$$\mathbf{Z}^M = [\mathbf{z}_{t_1}^{v,M}, \mathbf{z}_{t_1\to t_2}^{a,M}, \cdots, \mathbf{z}_{t_{L-1}}^{v,M}, \mathbf{z}_{t_{L-1}\to t_L}^{a,M}, \mathbf{z}_{t_L}^{v,M}]. \tag{4}$$

We then apply a GRU-based sequence encoder $E_\rho$ to further aggregate sequential information, and use ten prediction heads $\{H_{\xi_i}\}_{i=1}^{10}$ to predict the action toward each standard plane:

$$\mathbf{V} = E_\rho(\mathbf{Z}^M), \quad \mathbf{a}'_{t_L \to t_i} = H_{\xi_i}(\mathbf{v}_{t_L}), \tag{5}$$

where $\mathbf{V} = [\mathbf{v}_{t_1}, \ldots, \mathbf{v}_{t_L}]$ denotes the GRU outputs, and $\mathbf{v}_{t_L}$ corresponds to the current frame $\mathbf{I}_{t_L}$. Finally, the loss is calculated using the Smooth L1 Loss between the predicted action and the target $\mathbf{a}_{t_L \to t_{i\text{-th standard plane}}}$ as follows:

$$\mathcal{L} = \mathcal{L}_{\text{SmoothL1}}(\mathbf{a}_{t_L \to t_{i\text{-th standard plane}}}, \mathbf{a}'_{t_L \to t_{i\text{-th standard plane}}}). \tag{6}$$

In (6), translation and rotation are equally weighted; we normalize units in preprocessing (mm for translation, degrees for rotation) to match magnitudes.

## 3 Experiments

### 3.1 Datasets and Implementation Details

**Datasets.** Data were collected with a GE machine (*General Electric, USA*) equipped with a M5S probe (Section 2.1) under approval and supervision of the University Medical Ethics Committee. We use 284 scans for training and 72 for validation, with data from different subjects in each set.

**Model Architecture.** By default, the input sequence length is $L$=4 and the adapter bottleneck dimension (adapter dimension) is $r$=64, with ReLU activation. The core vision–action interaction module $S_\psi$ is a Transformer block with 4 attention heads and an MLP ratio of 2, using 1D sincos positional encoding. We initialize linear/conv weights with truncated normal ($\sigma$=0.02) and set biases to 0; LayerNorm weights are initialized to 1 and biases to 0. After the image encoder, image and action features are each projected to 128-d and concatenated as input to the GRU sequence encoder $E_\rho$ (input 256, hidden 128). Each prediction head $H_{\xi_i}$ is a two-layer MLP with GELU in between.

**Training Strategy.** We train with Adam (batch size 256) for 5 epochs, using a cosine-decayed learning rate from $1 \times 10^{-4}$ to $1 \times 10^{-6}$, and all experiments are performed on four A100 GPUs.

**Evaluation Metric.** We report trainable parameters and the Mean Absolute Error (MAE) between predicted $\mathbf{a}'$ and ground-truth $\mathbf{a}$, computed separately for translation and rotation:

$$\text{MAE of Trans.} = \frac{1}{3}\sum_{i=1}^{3}|\mathbf{a}_i - \mathbf{a}'_i|, \quad \text{MAE of Rot.} = \frac{1}{3}\sum_{i=4}^{6}|\mathbf{a}_i - \mathbf{a}'_i|. \tag{7}$$

### 3.2 Comparison with Baselines

We evaluated our method on ten standard views, as shown in Table 1. Our method outperforms baselines in both MAE and parameter efficiency. Although

Table 1: **Comparison with baseline methods in terms of translation and rotation dimensions.** Results include the number of trained parameters and the corresponding errors. "Avg." denotes the average Mean Absolute Error computed across ten standard planes. †indicates that these methods use the pre-trained encoder from DeiT.

| Setting | Method | Trained Params | Translation Avg. (mm) | Rotation Avg. (°) |
|---|---|---|---|---|
| Single-frame, ImageNet-pretrained | DeiT [23] | 22.60M | 8.43 | 8.84 |
| | DINOv2 [20] | 90.28M | 8.22 | 8.72 |
| Single-frame, Pre-trained on ultrasound data | BiomedCLIP [28] | 89.50M | 8.44 | 9.03 |
| | LVM-Med [17] | 89.44M | 8.56 | 8.90 |
| | US-MoCo [4] | 22.60M | 8.51 | 8.80 |
| | US-IJEPA [1] | 22.59M | 8.35 | 8.67 |
| | US-MAE [8] | 22.60M | 8.26 | 8.66 |
| | USFM [14] | 92.94M | 8.26 | 8.62 |
| | EchoCLIP [5] | 89.74M | 8.21 | 8.52 |
| Sequential, ImageNet-pretrained | US-GuideNet† [6] | 22.05M | 7.53 | 7.83 |
| | Decision-T† [3] | 22.27M | 7.38 | 7.88 |
| Sequential, Pre-trained on ultrasound data | BiomedCLIP | 86.15M | 7.29 | 7.80 |
| | **BiomedCLIP+Ours** | **3.94M** | **5.55** | **6.69** |
| | USFM | 87.04M | 7.15 | 7.81 |
| | **USFM+Ours** | **3.97M** | **5.35** | **6.71** |
| | EchoCLIP | 88.41M | 6.56 | 7.66 |
| | **EchoCLIP+Ours** | **2.61M** | **5.40** | **6.74** |

only averaged errors are reported, this improvement is consistent across all ten standard views. Single-frame models rely only on the current image and cannot capture structural variation, leading to poor performance. Baseline sequential models [3,6] typically fuse image and action features only in the prediction head, underusing structural cues in the sequence, and require full fine-tuning with high training cost. In contrast, our lightweight VA-Adapter is inserted into the image encoder, enabling progressive structure learning during feature extraction while improving both efficiency and accuracy.

We also compare with other common PEFT methods, including LoRA [9] and Prefix Tuning [16]. Since they are designed for transformer-based backbones, we evaluate them only on USFM and BiomedCLIP. As shown in Fig. 3, VA-Adapter performs best. This is likely because VA-Adapter explicitly models vision–action interactions to capture cardiac structure, whereas other PEFT methods mainly improve efficiency without enhancing structural understanding.

### 3.3 Ablation Study

**Vision-action Interaction Module.** We compare seven training baselines for EchoCLIP in Fig. 4. The vanilla adapter removes the vision–action interaction

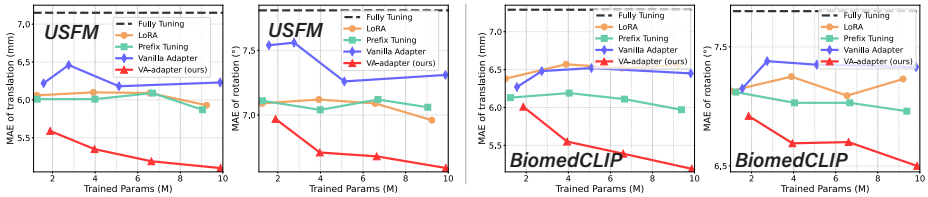


Fig. 3: Performance comparison of PEFT methods on USFM and BiomedCLIP.

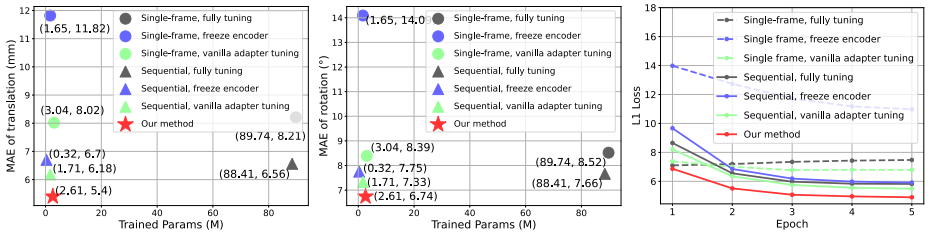


Fig. 4: Ablation on vision-action interaction module of the EchoCLIP model.

module while keeping other settings unchanged. With 2.61M trainable parameters, our method achieves the lowest MAE and the best parameter–accuracy trade-off. Compared with the vanilla adapter of similar scale, it reduces MAE by 12.6% / 8.0% (Trans./Rot.) with only 0.9M extra parameters, demonstrating that the performance gains from the vision-action interaction mechanism outweigh the marginal cost of extra parameters. Furthermore, it also demonstrates faster convergence, outperforming other methods after the first epoch.

**Adapter Dimension.** We study the effect of adapter dimension on performance (Fig. 5). With dimension $r$=8, our method introduces only 0.2M parameters (0.23% of full fine-tuning) yet reduces MAE by 11.9% / 8.2% (Trans./Rot.), whereas the vanilla Adapter with similar size achieves only 4.6% / 4.2%. This indicates that our VA-Adapter captures cardiac structural cues more effectively under tight parameter budgets. As $r$ increases from 8 to 128, our parameters grow by 31.7× while MAE further drops by 9.2% / 6.3%. In contrast, the vanilla adapter grows by 15.4× with only 0.2% / 2.0% additional MAE reduction, indicating that our interaction module uses added parameters more efficiently.

### 3.4 Visualization

We visualize the model outputs in Fig. 6. We apply the predicted action to obtain the resulting pose, then perform nearest-neighbor retrieval in the scan sequence and compare the retrieved plane with the target. Results show that the model guides the probe toward the target plane. The outputs are consistent across frames from the same probe position but different cardiac phases, indicating robustness to the cardiac cycle. Moreover, for non-standard or low-quality planes

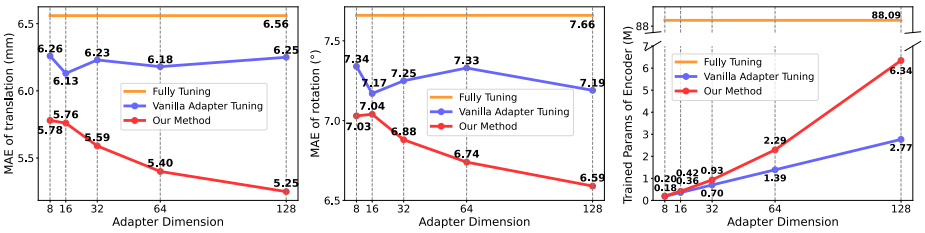


Fig. 5: Ablation study on adapter dimension of the EchoCLIP model.

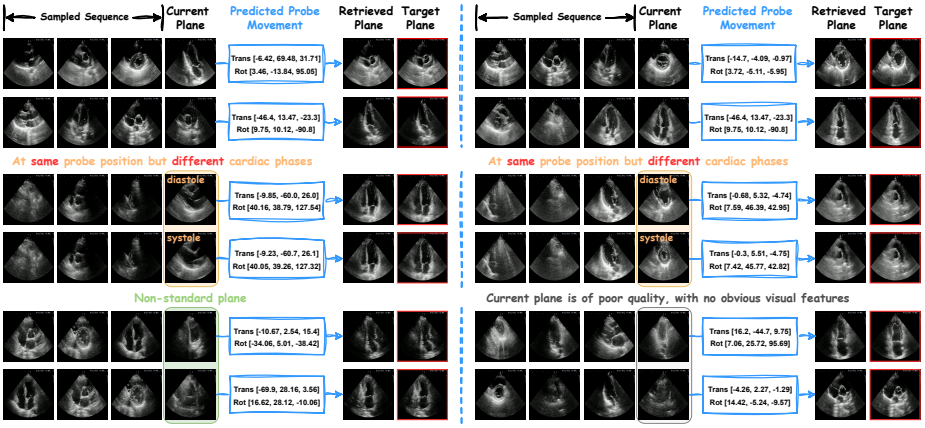


Fig. 6: Visualization of the model's prediction.

with weak visual cues, the sequence model can still infer correct actions from vision–action relationships, which single-frame models cannot.

### 3.5 Inference Real-time Analysis

We benchmark inference on RTX 3090. Across all backbones, a single sequence takes ∼10.0–11.1 ms without VA-Adapter and ∼10.5–11.8 ms with VA-Adapter, satisfying real-time requirements. The added latency is marginal, indicating VA-Adapter preserves deployment efficiency in time-sensitive ultrasound guidance.

## 4 Conclusion

In this paper, we propose VA-Adapter, a lightweight module that empowers ultrasound foundation models with probe guidance capability by modeling vision-action interactions during feature encoding. By combining basic knowledge from the foundation model with personalized 3D cardiac structures learned from the VA-Adapter, our method updates 95.4%–97.0% fewer parameters and reduces guidance error by 12.0%–25.2%, with extensive experiments validating superior

efficiency and performance over state-of-the-art baselines. This superior performance supports practical clinical scenarios, such as offering real-time assistance to junior sonographers and serving as the decision-making core for fully autonomous robotic ultrasound systems.

**Acknowledgments.** This work is supported in part by the National Key Research and Development Program of China under Grant 2024YFB4708200, and the Scientific Research Innovation Capability Support Project for Young Faculty under Grant ZYGXQNJSKYCXNLZCXM-I20, and the National Natural Science Foundation of China under Grant 62461160307.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023)
2. Bao, M., Wang, Y., Wei, X., Jia, B., Fan, X., Lu, D., Gu, Y., Cheng, J., Zhang, Y., Wang, C., et al.: Real-world visual navigation for cardiac ultrasound view planning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 317–326. Springer (2024)
3. Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., Mordatch, I.: Decision transformer: Reinforcement learning via sequence modeling. Advances in neural information processing systems **34**, 15084–15097 (2021)
4. Chen, X., Xie, S., He, K.: An empirical study of training self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9640–9649 (2021)
5. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine pp. 1–8 (2024)
6. Droste, R., Drukker, L., Papageorghiou, A.T., Noble, J.A.: 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. pp. 583–592. Springer (2020)
7. Ghorbani, A., Ouyang, D., Abid, A., He, B., Chen, J.H., Harrington, R.A., Liang, D.H., Ashley, E.A., Zou, J.Y.: Deep learning interpretation of echocardiograms. NPJ digital medicine **3**(1), 10 (2020)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: Lora: Low-rank adaptation of large language models. CoRR **abs/2106.09685** (2021)
10. Jiang, H., Li, M., Sun, Z., Jia, N., Sun, Y., Luo, S., Song, S., Huang, G.: Structure-aware world model for probe guidance via large-scale self-supervised pre-train. arXiv preprint arXiv:2406.19756 (2024)

11. Jiang, H., Sun, Z., Jia, N., Li, M., Sun, Y., Luo, S., Song, S., Huang, G.: Cardiac copilot: Automatic probe guidance for echocardiography with world model. arXiv preprint arXiv:2406.13165 (2024)
12. Jiang, H., Wang, T., Sun, Z., Wang, Y., Yue, Y., Sun, Y., Jia, N., Li, M., Luo, S., Song, S., et al.: Ultrasep: Sequence-aware pre-training for echocardiography probe movement guidance. Pattern Recognition p. 112600 (2025)
13. Jiang, H., Zhao, A., Yang, Q., Yan, X., Wang, T., Wang, Y., Jia, N., Wang, J., Wu, G., Yue, Y., et al.: Towards expert-level autonomous carotid ultrasonography with large-scale learning-based robotic system. Nature Communications **16**(1), 7893 (2025)
14. Jiao, J., Zhou, J., Li, X., Xia, M., Huang, Y., Huang, L., Wang, N., Zhang, X., Zhou, S., Wang, Y., et al.: Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Medical Image Analysis **96**, 103202 (2024)
15. Li, K., Li, A., Xu, Y., Xiong, H., Meng, M.Q.H.: 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)
16. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (2021)
17. MH Nguyen, D., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B., Swoboda, P., Ho, N., Albarqouni, S., Xie, P., 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)
18. Mitchell, C., Rahko, P.S., Blauwet, L.A., Canaday, B., Finstuen, J.A., Foster, M.C., Horton, K., Ogunyankin, K.O., Palma, R.A., Velazquez, E.J.: 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)
19. Narang, A., Bae, R., Hong, H., Thomas, Y., Surette, S., Cadieu, C., Chaudhry, A., Martin, R.P., McCarthy, P.M., Rubenson, D.S., 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)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C.P., Heidenreich, P.A., Harrington, R.A., Liang, D.H., Ashley, E.A., et al.: Video-based ai for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
22. Roth, G.A., Johnson, C., Abajobir, A., Abd-Allah, F., Abera, S.F., Abyu, G., Ahmed, M., Aksut, B., Alam, T., Alam, K., 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)
23. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
24. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)

25. Wang, T., Jiang, H., Wang, Y., Sun, Z., Yan, X., Li, X., Huang, G.: Ultrahit: A hierarchical transformer architecture for generalizable internal carotid artery robotic ultrasonography. arXiv preprint arXiv:2509.13832 (2025)
26. Yang, L., Zhang, R.Y., Wang, Y., Xie, X.: Mma: Multi-modal adapter for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23826–23837 (2024)
27. Yue, Y., Wang, Y., Jiang, H., Liu, P., Song, S., Huang, G.: Echoworld: Learning motion-aware world models for echocardiography probe guidance. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25993–26003 (2025)
28. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)