UltraStar: Semantic-Aware Star Graph Modeling for Echocardiography Navigation

Teng Wang[1*], Haojun Jiang[1*‡], Chenxi Li[1], Diwen Wang[2], Yihang Tang[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 Automation, Beijing Institute of Technology, Beijing, 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 critical for diagnosing cardiovascular diseases, yet the shortage of skilled sonographers hinders timely patient care, due to high operational difficulties. Consequently, research on automated probe navigation has significant clinical potential. To achieve robust navigation, it is essential to leverage historical scanning information, mimicking how experts rely on past feedback to adjust subsequent maneuvers. Practical scanning is an exploratory trial-and-error process that inherently generates noisy trajectories. However, existing methods typically model this history as a sequential chain, forcing models to overfit these noisy paths, leading to performance degradation on long sequences. In this paper, we propose UltraStar, which reformulates probe navigation from path regression to anchor-based global localization. By establishing a Star Graph, UltraStar treats historical keyframes as spatial anchors connected directly to the current view, explicitly modeling geometric constraints for precise positioning. We further enhance the Star Graph with a semantic-aware sampling strategy that actively selects the representative landmarks from massive history logs, reducing redundancy for accurate anchoring. Extensive experiments on a dataset with over 1.31 million samples demonstrate that UltraStar outperforms baselines and scales better with longer input lengths, revealing a more effective topology for history modeling under noisy exploration. Code is available at https://github.com/LeapLabTHU/UltraStar.

Keywords: Echocardiography · Star Graph Modeling · Semantic-aware Sampling · Probe Navigation.

1 Introduction

Cardiovascular diseases are the leading cause of death [19]. To assess cardiovascular health, echocardiography is an effective method [17], being both non-invasive and free from radiation. In practice, sonographers need to manually adjust and

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

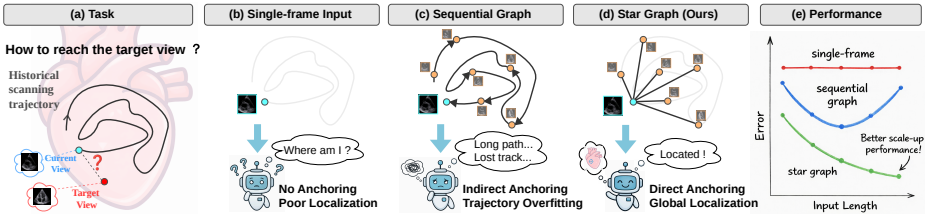

Fig. 1: Comparison of modeling paradigms. (a) The echocardiography navigation task. (b) Single-frame methods struggle with localization due to limited context. (c) Sequential Graph methods model history as a chain, forcing the model to overfit noisy exploration trajectories and degrading localization accuracy. (d) Our Star Graph breaks the chain, treating historical keyframes as global anchors and learning direct geometric constraints for precise localization. (e) Our method achieves lower error and scales better.

explore the probe between the ribs, continuously optimizing the angle based on previous exploration trajectory to approach the target view. This demanding process requires strong anatomical knowledge and proficient skills, contributing to the scarcity of trained professionals [21].

To address this issue, with the rapid adoption of AI techniques in medical imaging [6,15,16,23,24], researchers have begun leveraging AI to empower ultrasound probe navigation systems [2,4,8,13,14,18,22], aimed at achieving precise probe navigation to assist sonographers or drive autonomous robotic ultrasound systems. Several previous studies have demonstrated the potential of AI-driven guidance. For example, Narang et al. [18] proposed a commercial CNN-based echocardiography navigation system to guide nurses in acquiring standard views, but it is closed-source, limiting broader development. Bao et al. [2] proposed a convolutional model for rotational suggestions, but it is limited to A4C and A2C views, reducing its clinical scope. Droste et al. [8] introduced US-GuideNet, which uses a GRU to model historical image-motion sequences for fetal plane scanning. Jiang et al. [13] proposed UltraSeP, which leverages a sequence-aware pre-training paradigm to learn personalized cardiac structures. Wang et al. [20] developed UltraHiT, employing a hierarchical transformer to model history scanning sequences for autonomous internal carotid artery scanning. Additionally, Yue et al. [22] proposed EchoWorld, a motion-aware world modeling framework that pre-trains a cardiac world model to improve probe guidance.

As mentioned earlier, cardiac ultrasound scanning is typically a process of exploration followed by gradual convergence, where sonographers rely on prior exploration to guide subsequent maneuvers. Therefore, leveraging historical scanning data is indispensable for robust probe navigation, as single-frame input methods [11,12] lack sufficient context to resolve the ambiguity of the complex cardiac anatomy. While recent works adopt sequence modeling to capture this context, limitations still remain in the utilization of historical information. As illustrated in Fig. 1, cardiac ultrasound scanning is a complex trial-and-error

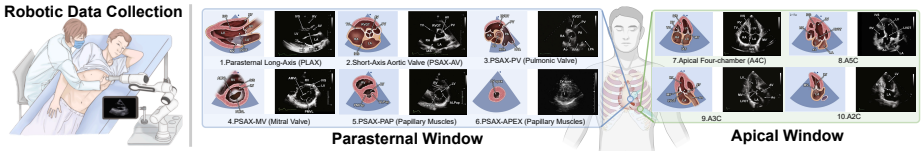

Fig. 2: Illustration of the dataset. The view images are sourced from [17].

exploration process. Consequently, the raw history contains substantial redundant or noisy motions that do not directly aid target acquisition. However, most existing sequential methods [8,13,20] model the history as a chain (Fig. 1(c)). This paradigm forces the model to overfit the noisy exploration path, paying excessive attention to step-by-step wandering motions. As a result, the geometric constraints needed for navigation are weakened, and performance often degrades as the sequence length grows. A key insight of our work is that the main value of history lies not in reconstructing the trajectory, but in global localization—helping the agent understand its current position relative to the anatomy. To this end, we propose UltraStar, a simple yet effective framework that shifts from chain-based modeling to a Star Graph topology (Fig. 1(d)). By treating historical keyframes as spatial anchors and directly connecting them to the current view, UltraStar learns the geometric constraints between the current state and historical landmarks, enabling robust localization and superior scalability. Furthermore, raw scanning history can be excessively long and sparse, making full-trajectory processing computationally infeasible. We therefore propose a semantic-aware sampling strategy that selects keyframe landmarks with high semantic divergence, constructing a compact yet informative map for the Star Graph. We validate our method on a large-scale dataset of 1.31 million samples, where UltraStar outperforms baselines and demonstrates superior scalability.

2 Dataset and Method

In this section, we first describe the echocardiography scanning dataset to provide a background about our task. Next, we introduce the Star Graph modeling paradigm. Lastly, we introduce the semantic-aware sampling strategy.

2.1 Echocardiography Scanning Dataset

In this work, we focus on six imaging planes in the parasternal window and four imaging planes in the apical window (Fig. 2 right). To support model training, we built a robotic acquisition system in which a sonographer operates a probe mounted on a robotic arm (Fig. 2 left), while the probe captures images and the system records its 6-DOF pose. The pose is represented by position (x, y, z) and orientation (Euler angles about x, y, z). Each scan yields a sequence $\{(\mathbf{I}_t, \mathbf{p}_t)\}_{t=1}^{T}$, where $\mathbf{I}_t$ is the image and $\mathbf{p}_t$ is the pose at time t, and the relative action $\mathbf{a}_{i\to j}$ between two poses $\mathbf{p}_i$ and $\mathbf{p}_j$ can be computed. During each scan, sonographers

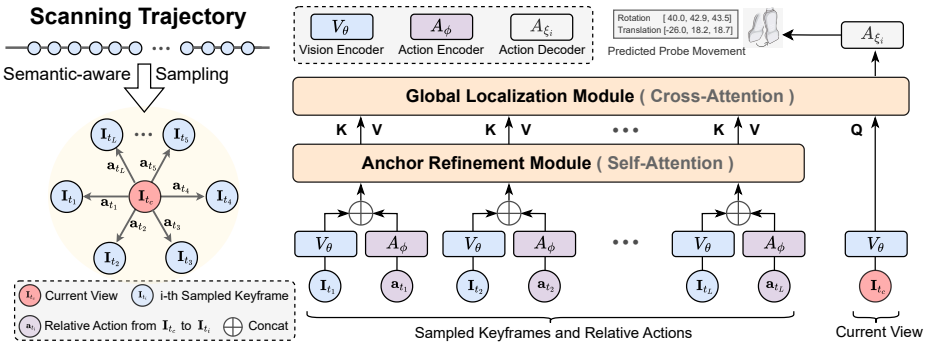

Fig. 3: Illustration of the Star Graph modeling paradigm.

annotate timestamps $\{t_i\}_{i=1}^{10}$ corresponding to the ten standard planes, from which the target poses $\mathbf{p}_{t_i}$ are obtained. For any other frame $(\mathbf{I}_{t_j}, \mathbf{p}_{t_j})$ with $j \neq i$, we compute the action $\mathbf{a}_{t_j \to t_i}$ to the target plane as supervision for training. Our dataset was collected by two expert sonographers from 178 adult patients, resulting in 356 scanning trajectories and 1.31M samples in total.

2.2 Star Graph Modeling Paradigm

The core motivation of our method is to utilize historical data for global localization rather than trajectory reconstruction. By explicitly modeling the geometric relationship between the current view and historical landmarks, the model can determine its precise location in the cardiac anatomy. In the following sections, we provide a detailed explanation of our approach, as shown in Fig. 3.

Input. Given the current image $\mathbf{I}_{t_c}$ from a scan, we first sample $L-1$ historical keyframes $\{\mathbf{I}_{t_1}, \ldots, \mathbf{I}_{t_{L-1}}\}$ using the strategy described in Sec. 2.3. Unlike sequential methods that model the path between adjacent historical frames, we construct a Star Graph topology. We compute the relative actions $\mathbf{a}_{t_c \to t_i} \in \mathbb{R}^6$ from the current view $\mathbf{I}_{t_c}$ to each sampled keyframe $\mathbf{I}_{t_i}$. This forms a set of spatial anchors, where each anchor consists of a visual appearance and its geometric position relative to the current view:

$$\mathcal{S} = \{(\mathbf{I}_{t_i}, \mathbf{a}_{t_c \to t_i}) \mid i = 1, \ldots, L-1\}. \tag{1}$$

Feature Encoding. We utilize a shared vision encoder V_θ (ViT [7]) to extract visual features. The current view $\mathbf{I}_{t_c}$ is encoded into a query feature $\mathbf{f}_c^v \in \mathbb{R}^C$. Similarly, each historical keyframe $\mathbf{I}_{t_i}$ is encoded into a visual feature $\mathbf{f}_i^v \in \mathbb{R}^C$. Simultaneously, the relative 6-DOF actions are processed by an action encoder A_ϕ to map the geometric constraints into the same feature space:

$$\mathbf{f}_c^v = V_\theta(\mathbf{I}_{t_c}), \quad \mathbf{f}_i^v = V_\theta(\mathbf{I}_{t_i}), \quad \mathbf{f}_i^a = A_\phi(\mathbf{a}_{t_c \to t_i}), \quad \mathbf{f} \in \mathbb{R}^C. \tag{2}$$

Star Graph Reasoning. The reasoning process consists of two stages: Anchor Refinement and Global Localization. First, to construct robust multimodal anchors, we concatenate the visual and action features for each sampled

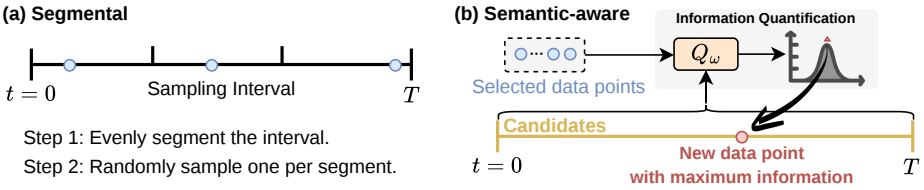

Fig. 4: Diagram of segmental sampling and idea of semantic-aware sampling.

keyframe. These raw anchors are then fed into an Anchor Refinement Module (implemented as a two-layer Self-Attention block) to model the correlations among historical landmarks and filter noise:

$$\mathbf{h}_i = \text{Concat}(\mathbf{f}_i^v, \mathbf{f}_i^a) \in \mathbb{R}^{2C}, \tag{3}$$

$$[\hat{\mathbf{h}}_1, \ldots, \hat{\mathbf{h}}_{L-1}] = \text{SelfAttn}([\mathbf{h}_1, \ldots, \mathbf{h}_{L-1}]). \tag{4}$$

Next, we employ a Global Localization Module. We project the refined anchors $\hat{\mathbf{h}}$ to dimension C as the Key and Value, and use the current view feature $\mathbf{f}_c^v$ as the Query for the Cross-Attention mechanism. This operation aggregates geometric evidence from the map to localize the current state:

$$\mathbf{m} = \text{CrossAttn}(Q = \mathbf{f}_c^v, K = \hat{\mathbf{h}}, V = \hat{\mathbf{h}}). \tag{5}$$

Prediction and Loss. Finally, the localized feature $\mathbf{m}$ is passed to task-specific action decoders A_{ξ_k} to predict the relative action $\mathbf{a}_{t_c \to \text{target}_k}$ required to reach the k-th standard plane. We adopt millimeters for translation and degrees for rotation. This unit choice brings both components to the same order of magnitude, allowing us to assign equal weights to translation and rotation errors during optimization. The model is trained using the Smooth L1 Loss:

$$\mathbf{a}' = A_{\xi_k}(\mathbf{m}), \quad \mathcal{L} = \mathcal{L}_{\text{SmoothL1}}(\mathbf{a}_{\text{gt}}, \mathbf{a}'). \tag{6}$$

2.3 Semantic-aware Sampling Strategy

The efficacy of our UltraStar framework relies on the quality of the spatial anchors used for map construction. To achieve precise global localization, the constructed Star Graph must be informationally dense: a set of redundant anchors provides weak geometric constraints, whereas a diverse map facilitates robust triangulation. Let L denote the number of nodes in the input graph. Our goal is to select $L-1$ informative historical anchors from the raw scanning history.

Segmental Sampling. As a representative baseline, segmental sampling divides the historical trajectory into $L-1$ equal temporal segments and randomly selects one point from each segment (see Fig. 4(a)). This strategy improves temporal coverage compared with unconstrained sampling by encouraging anchors to be distributed across the full exploration process. However, it remains content-agnostic and does not consider the actual visual information, often resulting in homogeneous anchors that fail to represent the global anatomy effectively.

Semantic-aware Strategy. To construct an informationally rich map for the Star Graph, we propose a semantic-aware strategy that enhances basic sampling by maximizing the semantic diversity of the anchors (see Fig. 4(b)). Specifically, we train a view classification model Q_ω to quantify the semantic content of each frame. For an image $\mathbf{I}$, Q_ω outputs a probability distribution $\mathbf{z} \in \mathbb{R}^{10}$ over 10 standard views. The sampling process is designed to minimize redundancy. When selecting a new anchor, we compute its semantic similarity against the current view $\mathbf{I}_t$ and all previously sampled anchors using cosine similarity on their distribution vectors $\mathbf{z}$. We sum these similarity scores to measure the redundancy of a candidate frame relative to the existing graph nodes. To ensure diversity while maintaining randomness, we identify the K candidates with the lowest total similarity (i.e., the most distinct frames) and randomly select one as the new anchor. This iterative process ensures that the resulting Star Graph covers the widest possible semantic range within the cardiac anatomy.

3 Experiments

3.1 Implementation Details

Dataset. The dataset was collected using a GE machine (*General Electric*) equipped with a M5S probe. We use 284 scans for training and 72 scans for validation, with a strict patient-level split. The validation set comprises individuals not encountered during training. The vision encoder V_θ and view classification model Q_ω were trained only on training scans to prevent data leakage. The data collection was approved by the University Medical Ethics Committee.

Model Architecture. We use a ViT-Small/16 model as the default vision encoder. The action encoder is a linear layer that maps the original 6-DOF actions to a 192-dimensional action space to match the vision feature dimension. The Anchor Refinement Module is implemented as a two-layer Self-Attention block with 4 attention heads. The Global Localization Module consists of a single Cross-Attention layer with 4 attention heads. The action decoder is a two-layer MLP with a GELU function in between. For the view classification model Q_ω, we adopt a ResNet-34 [10] trained on 123K annotated samples.

Training Details. The model is optimized with AdamW (batch size 128, learning rate 1×10^{-4}) for 5 epochs using a cosine decay schedule. The vision encoder V_θ loads I-JEPA [1] parameters pre-trained on our dataset and is frozen during training. For semantic-aware sampling, the candidate selection hyperparameter K is set to 128. All experiments are performed on 8 H20 GPUs.

Evaluation Metric. The metric used is the Mean Absolute Error (MAE) between the predicted probe movement action $\mathbf{a}^{'}$ and the ground truth $\mathbf{a_{gt}}$.

3.2 Results and Analysis

Comparison with Baseline. Table 1 presents the quantitative comparison with an input graph size of $L = 8$. To isolate the impact of modeling paradigms,

Table 1: Comparison with baselines using MAE for **translation (mm)** and **rotation (°)**. We report the average MAE over six standard views from Parasternal Window, four standard views from Apical Window (Fig. 2), and the overall average across all views. * indicates $p < 0.0001$.

Type	Method	Parasternal		Apical		Average	
		Trans.	Rot.	Trans.	Rot.	Trans.	Rot.
Single-frame	BioMedCLIP [24]	8.24	8.55	8.75	9.76	8.44	9.03
	LVM-Med [16]	8.35	8.46	8.86	9.56	8.56	8.90
	US-MoCo [5]	8.37	8.44	8.74	9.60	8.51	8.90
	US-IJEPA [1]	8.13	8.20	8.67	9.38	8.35	8.67
	US-MAE [9]	8.13	8.19	8.46	9.36	8.26	8.66
	USFM [15]	8.12	8.19	8.36	9.27	8.21	8.62
	EchoCLIP [6]	8.14	8.16	8.33	9.05	8.21	8.52
Sequential Graph	GRU (from US-GuideNet [8])	6.97	7.20	5.09	8.26	6.22	7.62
	Causal self-attn (from Decision-T [3])	7.04	7.09	4.80	8.18	6.14	7.53
	Non-causal self-attn (from UltraSeP [13])	6.45	6.70	4.39	7.78	5.63	7.13
FC Graph	Motion-aware attn (from EchoWorld [22])	5.60	6.09	4.00	7.83	4.96	6.79
Star Graph	**UltraStar (Ours)**	**5.30**	**5.67**	**3.60**	**7.50**	**4.62** (↓7%)	**6.40** (↓6%)

*

all non-single-frame methods share the same frozen vision encoder. To ensure a rigorous evaluation, we explicitly exclude frames in the scanning trajectory that are visually similar to the target view from the candidate pool, preventing potential data leakage. As observed, UltraStar achieves state-of-the-art performance across all metrics. Sequential Graph methods lag behind because the chain structure encourages overfitting to noisy exploration trajectories. The Fully Connected Graph also underperforms compared to our method; its dense connectivity introduces excessive noisy edges irrelevant to the current decision, thereby confusing the model. In contrast, Star Graph establishes direct geometric links between historical anchors and the current view, enabling precise localization. Furthermore, we analyze the impact of input length in Fig. 5 (left). As the input graph size increases, UltraStar demonstrates superior scalability, with a consistent reduction in navigation errors. This indicates that our method effectively leverages extended historical information to refine localization, whereas baselines saturate or degrade due to the noise introduced by longer trajectories. We further conduct an ablation study on sampling strategies, as shown in Fig. 5 (right). Our semantic-aware sampling strategy consistently yields the lowest errors across all modeling paradigms, confirming its universality and effectiveness in selecting high-quality keyframes.

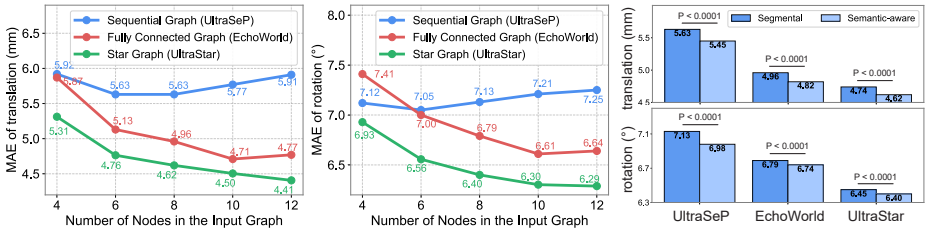

Fig. 5: **Left: Scalability analysis**. As the input graph size increases, our method demonstrates superior scalability with a consistent reduction in translation and rotation errors. **Right: Ablation study on sampling strategies**. We compare segmental sampling and semantic-aware sampling across different graph formulations, showing that semantic-aware sampling consistently yields lower translation and rotation errors.

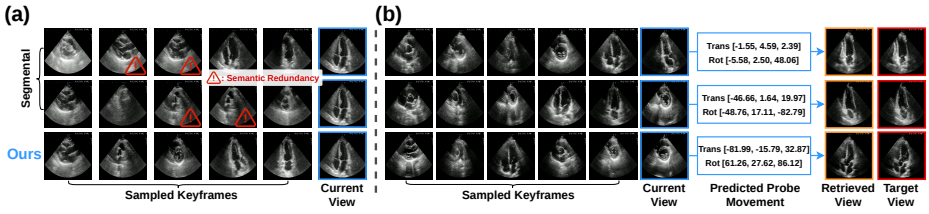

Fig. 6: (a) Visualization of the outcome of sampling strategies. (b) Visualization of the model prediction. Inputs include sampled keyframes, the current view, and the relative probe motions from the current view to each sampled keyframe.

Visualization. First, with the input graph size set to 6, Fig. 6(a) illustrates the outcomes of different sampling strategies. Segmental sampling exhibits distinct semantic redundancy, where multiple sampled keyframes convey repetitive visual information. In contrast, our semantic-aware strategy selects diverse landmarks, effectively maximizing the information entropy of the Star Graph. Second, to intuitively demonstrate the model's output, we calculate the final view pose based on the current view's pose and predicted probe movement action. Using this pose, we perform the nearest-neighbor retrieval from the same scan. We can observe from Fig. 6(b) that our model can output correct probe movement actions, bringing the probe closer to target views.

4 Conclusion

In this paper, we presented UltraStar, a framework that reformulates echocardiography probe navigation from path modeling to anchor-based global localization. By replacing the sequential chain with a Star Graph topology, UltraStar leverages historical frames as spatial anchors for geometric reasoning, establishing direct spatial constraints that ensure precise global localization. We further in-

troduce a semantic-aware sampling strategy to select representative landmarks, constructing a compact yet informative map from massive scanning logs. Experiments on a large-scale dataset show that UltraStar achieves state-of-the-art performance and scales better with longer history lengths, enabling precise and scalable ultrasound probe navigation. Beyond echocardiography, UltraStar offers a general anchor-based perspective for history modeling under noisy exploration. Future work could further explore the deployment of UltraStar on a robotic ultrasound platform to evaluate its closed-loop performance in real-world settings.

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, R., Yan, X., Lv, K., Huang, G., Li, Z., Li, X.: Ultradp: Generalizable carotid ultrasound scanning with force-aware diffusion policy. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 20074–20080 (2025)
5. 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)
6. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. Nature Medicine pp. 1–8 (2024)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
8. 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)
9. 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)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. 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. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 58–67 (2024)
12. 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. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 190–199 (2024)
13. 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)
14. 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)
15. 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)
16. 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)
17. 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)
18. 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)
19. 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)
20. 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)
21. Won, D., Walker, J., Horowitz, R., Bharadwaj, S., Carlton, E., Gabriel, H.: Sound the alarm: The sonographer shortage is echoing across healthcare. Journal of Ultrasound in Medicine (2024)

22. 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)
23. Yue, Y., Wang, Y., Tao, C., Liu, P., Song, S., Huang, G.: Chexworld: Exploring image world modeling for radiograph representation learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20778–20788 (2025)
24. 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)