# EchoCardMAE: Video Masked Auto-Encoders Customized for Echocardiography

Xuan Yang[1], Rui Xu[1,2,4(✉)], Xinchen Ye[1], Zhihui Wang[1], Miao Zhang[1], Yi Wang[1], Xin Fan[1,2], Hongkai Wang[3], Qingxiong Yue[4], Xiangjian He[5(✉)], and Yen-Wei Chen[6]

[1] DUT School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China
xurui@dlut.edu.cn
[2] DUT-RU Co-Research Center of Advanced ICT for Active Life, Dalian, China
[3] Faculty of Medicine, Dalian University of Technology, Dalian, China
[4] Central Hospital of Dalian University of Technology, Dalian, China
[5] School of Computer Science, University of Nottingham Ningbo China, Ningbo, China
Sean.He@nottingham.edu.cn
[6] College of Information Science and Engineering, Ritsumeikan University, Osaka, Japan

**Abstract.** Echocardiography, a vital cardiac imaging modality, faces challenges due to limited annotated data, impeding the application of deep learning. This paper introduces EchoCardMAE, a customized masked video autoencoder framework designed to leverage unlabeled echocardiography data and enhance performance across diverse cardiac tasks. EchoCardMAE addresses key challenges in echocardiogram analysis through three innovations built upon masked video modeling (MVM): (1) Key Area Masking, which concentrates feature learning on the diagnostically relevant sector of the image; (2) Temporal-Invariant Alignment Loss, promoting feature consistency across different clips of the same echocardiogram; and (3) Reconstruction Denoising, improving robustness to speckle noise inherent in echocardiography. We comprehensively evaluated EchoCardMAE on three public datasets, demonstrating state-of-the-art results in ejection fraction (EF) estimation, Myocardial infarction (MI) prediction, and cardiac segmentation. For example, on the EchoNet-Dynamic dataset, EchoCardMAE achieved an EF estimation MAE of 3.78 and a left ventricular segmentation mDice of 92.96, surpassing existing methods. The code is available at https://github.com/m1dsolo/EchoCardMAE.

**Keywords:** Echocardiography · Mask Video Modeling · Foundation Model

## 1 Introduction

Echocardiography is a widely used non-invasive cardiac imaging modality that provides valuable information about cardiac structure and function, and plays

a key role in various clinical applications, including disease diagnosis, treatment planning, and patient monitoring. Tasks in echocardiography, such as cardiac structure segmentation, disease identification, and EF estimation, have always been of great clinical significance and research value.

However, most of the current methods for echocardiography focus on specific tasks [4] [9] [11]. Although these methods have achieved encouraging results, they usually lack the ability to effectively utilize unlabeled data and generalization across different tasks, which limits the performance of the model. At present, some works have tried to train foundation models for ultrasound data [1] [7] [8], but they primarily focus on two-dimensional ultrasound images and do not leverage the temporal information inherent in echocardiography, which makes them unable to be directly applied to the field of echocardiography. Therefore, it is necessary to construct a foundation model that contains rich spatial-temporal information of echocardiography and can be easily applied for various heart-related tasks.

The most successful approach to training foundation models involves self-supervised learning [19] [5]. Masked modeling, in particular, has emerged as a powerful technique for learning general visual representations by randomly masking portions of the input data and training a model to reconstruct the missing content [2] [6]. In the video domain, masked video modeling (MVM) [16] [17] [20] [21] has shown promise. This paper explores the application of MVM to echocardiography. However, directly applying existing MVM methods to echocardiography is not ideal due to the unique characteristics of echocardiograms. Echocardiograms present several challenges compared to natural images: (1) First, the most relevant diagnostic information in echocardiograms is concentrated within the central, sector-shaped region. Directly applying MVM methods designed for natural images can lead to the reconstruction of large amounts of irrelevant background information, potentially biasing the model and wasting computational resources. (2) Echocardiograms typically encompass multiple cardiac cycles. In downstream tasks like identification or regression, it's often necessary to randomly select a video clip as input to the network, introducing variability into the experimental results. A lack of robustness to this variability can lead to inconsistent performance in downstream tasks. (3) Echocardiograms are inherently corrupted by speckle noise [23], which obscures subtle diagnostic features. In the context of MVM, speckle noise can mislead the reconstruction process, causing the model to prioritize reconstructing noise patterns rather than meaningful anatomical structures and motion. This ultimately hinders the learning of robust representations.

To address these challenges, this paper proposes **EchoCardMAE**, an MVM framework specifically customized for echocardiography. EchoCardMAE incorporates three key innovations: (1) To improve training efficiency and focus on diagnostically relevant regions, we implement key area masking, concentrating the masking process on the central, sector-shaped region of the echocardiogram. A learnable background token is also introduced to preserve positional encoding and background context. (2) To effectively leverage temporal information and re-

duce the sensitivity to clip selection, we propose a temporal-invariant alignment loss. This loss encourages feature consistency across different clips of the same echocardiogram. (3) To mitigate the effect of speckle noise, we employ a reconstruction denoising strategy. Instead of reconstructing the original noisy patches, the model is trained to reconstruct denoised versions, enhancing robustness to noise and improving representation quality.

In summary, the main contributions of this paper are:

1. We introduce EchoCardMAE, a novel masked video autoencoder framework, to address the unique challenges of echocardiogram analysis. EchoCardMAE establishes a foundation model suitable for diverse tasks, as demonstrated by extensive experiments on three public datasets, achieving state-of-the-art results on multiple tasks including EF estimation, left ventricular segmentation, and myocardial infarction prediction.
2. We introduce key area masking, targeting the sector-shaped region of echocardiograms where diagnostic information is concentrated, to improve training efficiency by focusing computation on the most important areas.
3. We propose a temporal-invariant alignment loss that enforces feature consistency across different clips, reducing sensitivity to clip selection and improving temporal representation learning.
4. We employ a reconstruction denoising strategy to mitigate the impact of speckle noise, enhancing the model's robustness to noise and improving representation quality.

## 2 Method

### 2.1 EchoCardMAE Training Pipeline

EchoCardMAE leverages MVM to address the challenges of echocardiogram analysis. The overall training pipeline is illustrated in Figure 1.

Given an unlabeled echocardiography dataset $D := \{X_i\}_{i=1}^{N}$, where each $X_i \in \mathbb{R}^{T\times H\times W\times 3}$ represents an echocardiography video, our objective is to train an encoder that learns generalizable representations suitable for various downstream tasks.

For each video $X_i$, we generate two clips, $X_{i1}$ and $X_{i2}$, by sampling with different starting frames. We then apply **key area masking** to the central sector region and input the unmasked patches to the encoder. Given the repetitive nature of heartbeats in echocardiograms, the features $F_{i1}$ and $F_{i2}$ extracted from $X_{i1}$ and $X_{i2}$ should exhibit temporal consistency. To enforce this, we employ a **temporal-invariant alignment loss**, which reduces sensitivity to clip selection. For reconstruction, the encoder's output is combined with learnable masked region embeddings and a background token to maintain contextual information. These combined features are then processed by the decoder. Finally, a **reconstruction denoising strategy** is applied to enhance the model's robustness to noise.

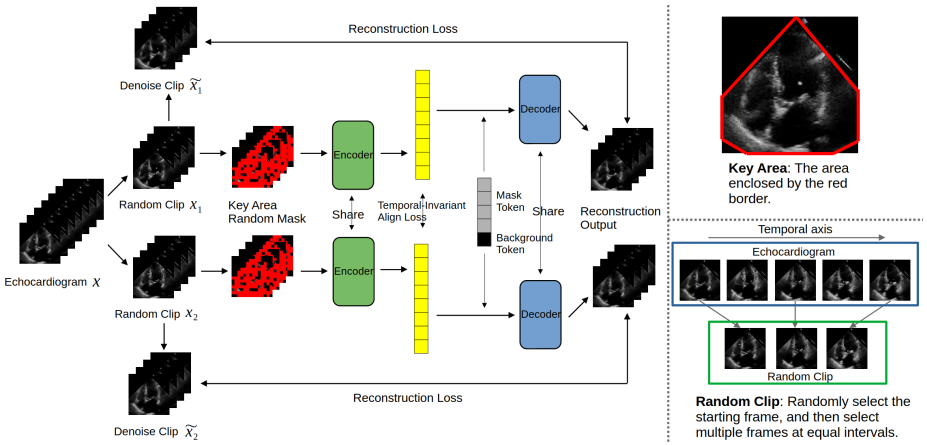


**Fig. 1.** Overall training pipeline of EchoCardMAE

The model is trained by minimizing the total loss, which is a weighted sum of the L2 reconstruction loss ($\mathcal{L}$rec) and the temporal-invariant alignment loss ($\mathcal{L}$align), as shown in the following equation:

$$\mathcal{L}\text{total} = \mathcal{L}\text{rec} + \lambda \mathcal{L}_{\text{align}}, \tag{1}$$

where $\lambda$ is a weighting factor that balances the two loss terms. In our experiments, $\lambda$ is set to 0.2.

The following sections provide a detailed description of these three key innovations.

### 2.2 Key Area Masking

Based on the prior knowledge that diagnostically relevant information is concentrated in the central sector of echocardiograms, we introduce key area masking. This method selectively masks only the central sector, while a learnable background token preserves positional encoding and background context. By focusing feature extraction on the diagnostically relevant sector and minimizing computation on irrelevant regions, we significantly improve training efficiency. We explored various masking strategies, including tube masking [17] (where entire temporal tubes are masked), blockwise masking [2] (where contiguous blocks of patches are masked), and random masking [6] (where patches are randomly selected). Random masking yielded the best performance for echocardiograms. We hypothesize that, at high masking ratios, random masking avoids overly difficult reconstruction tasks, leading to a smoother and more stable training process compared to the more structured tube or blockwise masking approaches.

### 2.3 Temporal-Invariant Alignment Loss

Echocardiograms inherently capture multiple cardiac cycles. To leverage this temporal information and reduce sensitivity to clip selection [15], we introduce a temporal-invariant alignment loss. This loss promotes feature consistency across different clips of the same echocardiogram, leading to more robust learned representations. Specifically, we utilize the InfoNCE loss [14], where $B$ is the batch size, $\tau$ is a temperature hyper-parameter set to 0.1 and $F_{i1}$ and $F_{i2}$ represent the feature embeddings extracted from two different clips of the same i-th echocardiogram in the batch.

$$\mathcal{L}_{\text{align}} = -\log\left(\frac{\exp(\mathbf{F_{i1}} \cdot \mathbf{F_{i2}}/\tau)}{\sum_{\substack{j=1 \\ j\neq i}}^{B} \exp(\mathbf{F_{i1}} \cdot \mathbf{F_{j2}}/\tau)}\right) \tag{2}$$

### 2.4 Reconstruction Denoising Strategy

Echocardiograms are inherently corrupted by noise. To address this, we enhance the model's robustness by modifying the reconstruction target during training. Instead of reconstructing the original noisy patches, we train the model to reconstruct denoised versions. Unlike the approach in [8], which need to denoise the input before feeding it to the model, our method eliminates the need for any pre-processing denoising during downstream inference, preserving real-time performance. We evaluated different denoising techniques, including Gaussian blur, median blur, and mean blur, and found that median blur yielded the best results for echocardiograms, possibly because its ability to effectively remove speckle noise in echocardiography, without overly blurring important structural details.

## 3 Experiments

### 3.1 Settings

All experiments were conducted using a single NVIDIA TITAN RTX (24GB) GPU and the PyTorch 2.5.1 framework. Following prior work [15], all data were resized to a resolution of 112x112 pixels. To capture the temporal dynamics of the heart, we randomly selected a starting frame for each echocardiography video and then sampled every 4th frame, resulting in a clip of 16 frames as input to the network. We chose ViT-S/16 [18] as the backbone encoder, as it offers a favorable balance between computational efficiency and performance, making it suitable for resource-constrained environments.

**Foundation model training**:

We performed foundation model training on the entirety of the training set from EchoNet-Dynamic, the largest publicly available echocardiography dataset (7465 echocardiography videos). Leveraging the insights from our masking strategy analysis, we employed random masking with a high mask ratio of 0.75,

promotes smoother and more stable training by avoiding overly difficult reconstruction tasks. We adopted AdamW as the optimizer with a weight decay of 0.5. Using L2 as the loss function, we trained the model with a batch size of 64 for 1600 epochs, with a learning rate of 1e-4.

**Fine-tuning**: The encoder trained with EchoCardMAE served as the foundation model for fine-tuning on three publicly available datasets. On EchoNet-Dynamic [15], we evaluated performance on both EF estimation and left ventricle segmentation tasks. On CAMUS [10], we evaluated performance on EF estimation and segmentation of the left ventricle, myocardium, and left atrium. On HMC-QU [3], we evaluated performance on MI prediction.

### 3.2 Datasets

**EchoNet-Dynamic** [15] is the largest publicly available echocardiogram dataset, comprising 10,030 A4C echocardiogram videos. Each video contains multiple cardiac cycles, with a frame size of 112x112 pixels. Each echocardiography video is accompanied by corresponding EF and left ventricle segmentation labels. We adhered to the data split defined in [15], with an approximate 6:1:1 ratio for training, validation, and testing sets.

**CAMUS** [10] consists of 500 cardiac cycle videos (A2C and A4C views) with EF values and segmentation results for the left ventricle, myocardium, and left atrium. For a fair EF estimation comparison, we used only the A4C view and the ten-fold cross-validation split from [12]. For left ventricle segmentation, we used the official challenge test set [10] and split the remaining data into training and validation sets (9:1).

**HMC-QU** [3] contains 109 A4C videos (72 with myocardial infarction and 37 normal). We follow the stratified 5-fold split provided by [13].

## 4 Results

### 4.1 Ejection Fraction Regression

This section presents a comparative analysis of EchoCardMAE against state-of-the-art methods for EF estimation. On the EchoNet-Dynamic [15] dataset, the largest publicly available echocardiograpy dataset, Table 1 shows that EchoCardMAE achieves superior performance, with a MAE of 3.78, RMSE of 4.94, and R2 of 0.84. We further assessed the transfer learning capabilities of EchoCardMAE on the CAMUS dataset, evaluating performance on both Good & Medium (G&M) and Poor (P) quality videos, following the experimental settings of [12]. As shown in Table 2, EchoCardMAE demonstrates improved transfer generalization compared to existing techniques.

### 4.2 Cardiac Segmentation

We further validated EchoCardMAE on cardiac segmentation. On EchoNet-Dynamic, we achieved state-of-the-art left ventricle (LV) segmentation, with a

**Table 1.** Performance Comparison on EF Estimation

| Method | MAE | RMSE | R2 |
|---|---|---|---|
| EchoNet [15] | 4.05 | 5.32 | 0.81 |
| EchoCoTr [13] | 3.95 | 5.17 | 0.82 |
| CoReEcho [12] | 3.90 | 5.13 | 0.82 |
| EchoMEN [9] | 3.93 | - | - |
| CardiacNet [22] | 3.83 | - | - |
| EchoCardMAE | **3.78** | **4.94** | **0.84** |

**Table 2.** Comparison EF regression(A4C view) with 10-fold CV

| Methods | G&M qual. | | P qual. | |
|---|---|---|---|---|
| | corr | MAE | corr | MAE |
| EchoCoTr [13] | 0.799 | 5.33 | 0.599 | 7.49 |
| CoReEcho [12] | 0.807 | 5.29 | 0.693 | 6.81 |
| EchoCardMAE | **0.846** | **5.06** | **0.717** | **5.41** |

mDice score of 92.96 and mIoU of 86.85 (Table 3). While performance on the CAMUS dataset (segmenting the left ventricle, myocardium, and left atrium) may be limited by the lower resolution of our training data, EchoCardMAE still demonstrates strong generalization, suggesting a robust foundation for echocardiography analysis.

**Table 3.** Comparison cardiac segmentation with SOTA methods on the test set

| Methods | EchoNet-Dynamic | | CAMUS | |
|---|---|---|---|---|
| | mDice | mIoU | mDice | mIoU |
| SAMUS [11] | 91.79 | 84.32 | 91.11 | 83.94 |
| MemSAM [4] | 92.78 | 85.89 | **93.31** | **87.61** |
| EchoCardMAE | **92.96** | **86.85** | 92.90 | 87.07 |

### 4.3 Myocardial infarction Prediction

We further assessed the classification capabilities of our model by performing a MI prediction task on the HMC-QU dataset. The results, presented in Table 4, demonstrate that our method achieves a superior F1 score, specificity, and accuracy compared to other methods.

**Table 4.** Comparison of MI classification (A4C view) with stratified 5-fold CV.

| Method | Sens. | Spec. | Prec. | F1 | Acc. |
|---|---|---|---|---|---|
| EchoCoTr [13] | **98.57** | 72.50 | 88.37 | 92.96 | 89.77 |
| CoReEcho [12] | 92.95 | 86.07 | **93.55** | 92.91 | 90.64 |
| EchoCardMAE | 95.71 | **86.79** | 93.54 | **94.51** | **92.63** |

### 4.4 Ablation Study

**Table 5.** Ablation study of three improvements of EchoCardMAE on EchoNet-Dynamic

| Key Area Mask | Denoising Rec | Align Loss | MAE |
|---|---|---|---|
| × | × | × | 3.8894 |
| ✓ | × | × | 3.8205 |
| ✓ | ✓ | × | 3.7991 |
| ✓ | ✓ | ✓ | **3.7768** |

**Table 6.** Comparison of different denoising methods on EchoNet-Dynamic

| Denoising Method | MAE |
|---|---|
| Mean Blur | 3.8062 |
| Gaussian Blur | 3.7839 |
| Median Blur | **3.7768** |

Table 5 shows the effectiveness of our three VideoMAE-based improvements customized for echocardiography analysis. Table 6 shows the effectiveness of different denoising methods for echocardiography analysis. Figure 2 illustrates that employing a temporal-invariant alignment loss reduces the impact of clip selection on EF estimation.

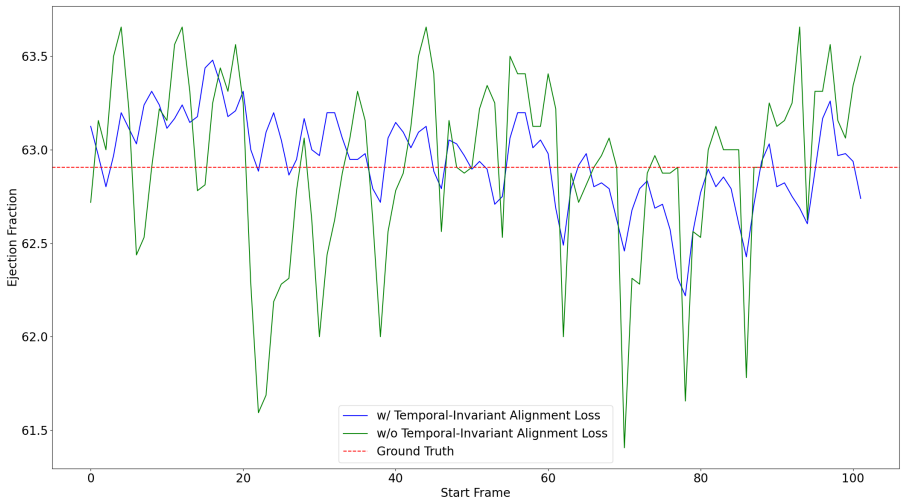


**Fig. 2.** EF variation across different clip start frames is smaller with the temporal-invariant alignment loss. This indicates that the model's EF estimation becomes less sensitive to the specific starting frame of the input clip when the temporal-invariant alignment loss is applied, leading to more robust and consistent performance.

## 5 Conclusion

EchoCardMAE, a customized masked video autoencoder, addresses key challenges in echocardiography through key region masking, temporal-invariant alignment and reconstruction denoising strategy. Evaluations on three datasets demonstrate state-of-the-art performance, establishing EchoCardMAE as a robust foundation for cardiac applications. While the relatively long training time of the foundation model (approximately 2 days) and the potential for improved performance on high-resolution images represent limitations, future work will address these areas to enhance clinical applicability.

**Acknowledgments.** This work was partially supported by the Fundamental Research Funds for the Central Universities of China under grant DUT23YG225, the funding of Dalian Key Laboratory of Digital Medicine for Critical Diseases, the Yongjiang Technology Innovation Project (2022A-097-G), National Natural Science Foundation of China talent grant (UNNC ID: B0166), and the Grant in Aid for Scientific Research from the Japanese Ministry for Education, Science, Culture and Sports (MEXT) under the Grant Nos. 20KK0234.

**Disclosure of Interests.** All authors declare that they have no conflicts of interest.

## References

1. Amadou, A.A., Zhang, Y., Piat, S., Klein, P., Schmuecking, I., Passerini, T., Sharma, P.: Echoapex: A general-purpose vision foundation model for echocardiography. arXiv preprint arXiv:2410.11092 (2024)
2. Bao, H., Dong, L., Piao, S., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021)
3. Degerli, A., Kiranyaz, S., Hamid, T., Mazhar, R., Gabbouj, M.: Early myocardial infarction detection over multi-view echocardiography. Biomedical Signal Processing and Control **87**, 105448 (2024)
4. Deng, X., Wu, H., Zeng, R., Qin, J.: Memsam: Taming segment anything model for echocardiography video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9622–9631 (2024)
5. Fadnavis, S., Parmar, C., Emaminejad, N., Ulloa Cerna, A., Malik, A., Selej, M., Mansi, T., Dunnmon, P., Yardibi, T., Standish, K., et al.: Echofm: A view-independent echocardiogram model for the detection of pulmonary hypertension. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–263. Springer (2024)
6. 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)
7. 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)
8. Kang, Q., Gao, J., Li, K., Lao, Q.: Deblurring masked autoencoder is better recipe for ultrasound image recognition. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 352–362. Springer (2023)

9. Lai, S., Zhao, M., Zhao, Z., Chang, S., Yuan, X., Liu, H., Zhang, Q., Meng, G.: Echomen: Combating data imbalance in ejection fraction regression via multi-expert network. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 624–633. Springer (2024)
10. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
11. Lin, X., Xiang, Y., Yu, L., Yan, Z.: Beyond adapting sam: Towards end-to-end ultrasound image segmentation via auto prompting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 24–34. Springer (2024)
12. Maani, F.A., Saeed, N., Matsun, A., Yaqub, M.: Coreecho: Continuous representation learning for 2d+ time echocardiography analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 591–601. Springer (2024)
13. Muhtaseb, R., Yaqub, M.: Echocotr: Estimation of the left ventricular ejection fraction from spatiotemporal echocardiography. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 370–379. Springer (2022)
14. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
15. 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)
16. Pei, G., Chen, T., Jiang, X., Liu, H., Sun, Z., Yao, Y.: Videomac: Video masked autoencoders meet convnets. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22733–22743 (2024)
17. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
18. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
19. Vukadinovic, M., Tang, X., Yuan, N., Cheng, P., Li, D., Cheng, S., He, B., Ouyang, D.: Echoprime: A multi-video view-informed vision-language model for comprehensive echocardiography interpretation. arXiv preprint arXiv:2410.09704 (2024)
20. Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., Qiao, Y.: Videomae v2: Scaling video masked autoencoders with dual masking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14549–14560 (2023)
21. Wang, R., Chen, D., Wu, Z., Chen, Y., Dai, X., Liu, M., Jiang, Y.G., Zhou, L., Yuan, L.: Bevt: Bert pretraining of video transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14733–14743 (2022)
22. Yang, J., Lin, Y., Pu, B., Guo, J., Xu, X., Li, X.: Cardiacnet: Learning to reconstruct abnormalities for cardiac disease assessment from echocardiogram videos. In: European Conference on Computer Vision. pp. 293–311. Springer (2024)
23. Yu, Y., Acton, S.T.: Speckle reducing anisotropic diffusion. IEEE Transactions on image processing **11**(11), 1260–1270 (2002)