


# EchoViewCLIP: Advancing Video Quality Control through High-performance View Recognition of Echocardiography

Shanshan Song<sup>1</sup>, Yi Qin<sup>1</sup>, Honglong Yang<sup>1</sup>, Taoran Huang<sup>2</sup>, Hongwen Fei<sup>2</sup>,  
and Xiaomeng Li<sup>1</sup>

<sup>1</sup> Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

<sup>2</sup> Guangdong Cardiovascular Institute, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou, Guangdong Province, China

 eexmli@ust.hk

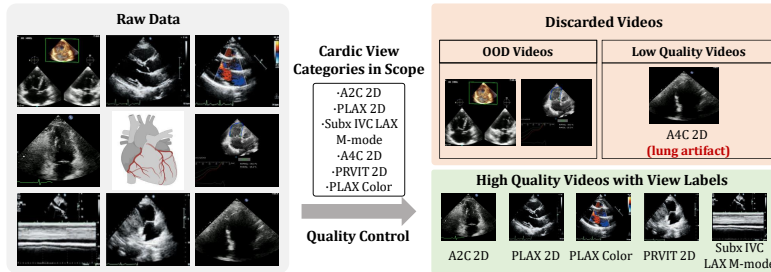
**Abstract.** Echocardiography is a critical imaging technique for diagnosing cardiac diseases, requiring accurate view recognition to support clinical analysis. Despite advancements in deep learning for automating this task, existing models face two major limitations: they support only a limited number of cardiac views, insufficient for complex cardiac diseases, and they inadequately handle out-of-distribution (OOD) samples, often misclassifying them into generic categories. To address these issues, we present EchoViewCLIP, a novel framework for fine-grained cardiac view recognition and OOD detection. Built on our collected large-scale dataset annotated with 38 standard views and OOD data, EchoViewCLIP integrates a Temporal-informed Multi-Instance Learning (TML) module to preserve temporal information and identify key frames, along with a Negation Semantic-Enhanced (NSE) Detector to effectively reject OOD views. Additionally, we introduce a quality assessment branch to evaluate the quality of detected in-distribution (ID) views, enhancing the reliability of echocardiographic analysis. Our model achieves 96.1% accuracy across 38 view recognition tasks. The code is available at <https://github.com/xmed-lab/EchoViewCLIP>.

**Keywords:** Echocardiography · View recognition · Quality screening.

## 1 Introduction

Echocardiography is a widely used cardiac imaging technique [2,25], providing key diagnostic evidence and guiding clinical decision-making [1,4,22,26]. Accurate cardiac diagnosis requires pathological findings from multiple cardiac views [21,24,28]. Therefore, precise view classification in echocardiography is crucial for enabling downstream analysis models to incorporate critical view-specific information, ensuring clinically meaningful and aligned interpretations.

Recent deep learning-based methods for echocardiography view classification can be categorized into image-based and video-based approaches. Image-based



**Fig. 1.** An example demonstrating the quality control process. **Raw data:** Original data collected by echocardiologists. **OOD Videos:** Videos that do not belong to any of the defined cardiac view classes. **Low Quality Videos:** Videos that fall within the defined view categories but exhibit insufficient quality to support reliable diagnosis. Both OOD and low-quality videos should be discarded. **High Quality Videos with View Labels:** The final high-quality data obtained through the quality control process readily support accurate and reliable echocardiographic analysis.

models treat video frames independently, fine-tuning pre-trained classifiers for view prediction [10,12,16,27,6,7,11,15]. However, they struggle with noisy or ambiguous frames, particularly for views that require spatio-temporal analysis across the entire video. This frame-level dependency limits their robustness and applicability to dynamic views. Video-based methods [18,20,14] could analyze full input videos to capture temporal relationships, which is better than image-based methods. Among video-based models, ViFi-CLIP [14] achieves advanced video classification performance by averaging pooled frame-level features and finetuned on contrastive language-image pre-training model (CLIP) [13]. However, its simplistic aggregation struggles to effectively distinguish similar inter-class features and varying intra-class characteristics in echocardiography videos, leading to suboptimal accuracy in multi-view recognition. Despite previous advancements, existing view classification models still encounter two key limitations. First, they support only a limited number of view classes, with most models restricted to a maximum of 23 standardized categories [27]. Comprehensive diagnostic assessments, particularly in complex cardiac conditions, necessitate the analysis of over 30 standardized views to adequately capture subtle anatomical variations [8]. Second, existing models struggle to handle OOD videos that deviate from predefined view categories (see Fig. 1). These models often misclassify them into arbitrary classes or assign them to a generic "Others" class. This introduces noise during training and undermines model reliability in real-world scenarios, where non-standard views are frequently encountered.

Our primary objective is to develop a high-performance view recognition model that supports a comprehensive range of cardiac view categories and enables quality review in real-world clinical settings as shown in Fig. 1. To this end, we first construct a large-scale echocardiography dataset annotated with 38 standard cardiac views and diverse OOD instances, meticulously labeled by nine experienced cardiologists. This dataset serves as a clinically grounded and bench-

mark for training and evaluating view classification models in alignment with practical clinical scenarios. Building upon this dataset, we introduce EchoViewCLIP, a novel CLIP-based framework that leverages both positive and negative semantics to perform 38-view ID recognition and OOD detection. The visual encoder comprises two visual experts to learn positive and negative semantics independently, each integrated with TML to preserve temporal dynamics and identify key frames critical for precise view classification. Leveraging negation semantics, we further develop the NSE OOD Detector, which enables the model to reject OOD views rather than misclassify them into ID categories. Moreover, the integration of TML and NSE enhances the model’s ability to perform accurate quality assessments on ID samples that belong to predefined views. This emergent capability facilitates more robust echocardiography data standardization, positioning EchoViewCLIP as a comprehensive tool for advancing clinical workflows in cardiac imaging. We summarize our contributions as follows:

- For accurate view classification, we propose the TML module, which preserves temporal dynamics and selects key frames for precise recognition. To ensure robust OOD detection, we develop the NSE module, which uses semantic contrast to reject OOD samples rather than misclassifying them.
- We extend EchoViewCLIP to include for ID quality assessment, further enhancing its application for comprehensive quality screening.
- Experimental results on our large dataset demonstrate EchoViewCLIP’s SOTA performance in both standard view classification and OOD detection.

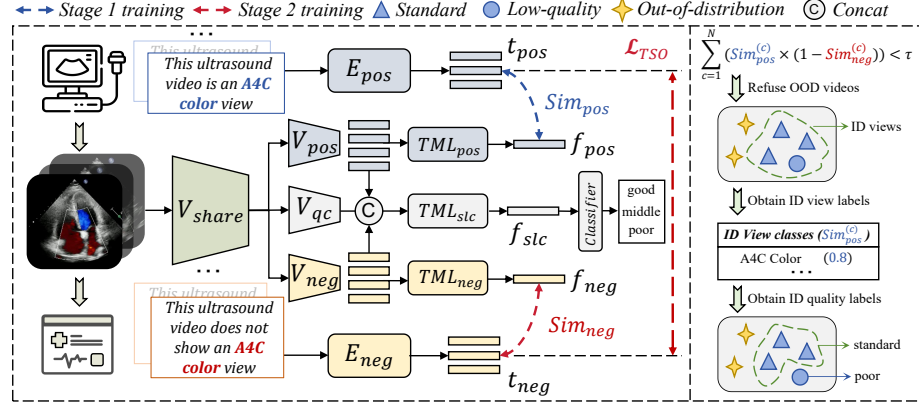
## 2 Methodology

**Overview.** EchoViewCLIP contains three core components, namely Standard View Classifier (SVC) with TML, NSE OOD Detector, and Quality Reviewer (QR), as illustrated in Fig. 2. Given an echocardiography video as input, the SVC and NSE jointly determine the video’s domain attribute and view category. If the input belongs to the ID category, the QR integrates both positive and negative semantics to assess the video’s quality.

### 2.1 Standard View Classifier with TML

The main goal of SVC is accurately classifying input echocardiography videos into the corresponding classes. We introduce temporal-informed multi-instance learning (TML) to preserve the temporal dependency among video frames while learning frame-specific feature weights. TML treats multiple frames as a set, allowing the model to focus on key frames crucial for view recognition. Specifically, we input an ultrasound video  $x \in \mathbb{X}^{T \times 3 \times 224 \times 224}$  of  $T$  frames through a visual encoder ( $V_{share}$  and  $V_{pos}$  in Fig. 2) to obtain the frame-wise visual representations  $h_{pos} = \{h_1, h_2, \dots, h_T\}$ . Then, the visual representations are used to calculate the attention weights  $w_t$  for each frame  $h_t$ :

$$w_t = \frac{\exp\{q^\top \tanh(Mh_t^\top)\}}{\sum_{j=1}^T \exp\{q^\top \tanh(Mh_j^\top)\}} \in [0, 1], \quad (1)$$



**Fig. 2.** Overall framework of our EchoViewCLIP. Given an input video, our model generates three outputs. First, by calculating  $sim_{pos}$  and  $sim_{neg}$ , the model determines whether the input is an OOD sample based on the threshold  $\tau$ . Then, if the input is an ID sample, the model predicts its corresponding cardiac view category according to  $sim_{pos}$ . Finally, based on the predicted view category, our model further evaluates the sample's quality label using our quality assessment branch.

where  $q$  and  $M$  are learnable parameters that determine the importance of each frame in the context of the entire video. The final video-level representation  $f_{pos}$  is then derived by combining the weighted frame-level features  $k_{pos}$  and the overall temporal semantic features  $t_{pos}$ , as follows:

$$f_{pos} = P(C(k_{pos}, t_{pos})), \text{ s.t. } k_{pos} = \sum_{t=1}^T w_t h_t, t_{pos} = \frac{1}{T} \sum_{t=1}^T h_t, \quad (2)$$

where  $C(\cdot)$  means concatenation and  $P(\cdot)$  is a linear projector.

We treat the view labels as structured text following ViFi-CLIP, and obtain the text representation set  $\mathcal{T}_{pos} = \{t_{pos}^1, t_{pos}^2, \dots, t_{pos}^N\}$ , where  $N$  is the number of standard views. The output is made by calculating the cosine similarity between the video-level representation  $f_{pos}$  and the text representation  $t_{label}$  as follows:

$$\hat{y} = \arg \max_i \text{sim}(f_{pos}, t_{pos}^i), \text{ s.t. } \text{sim}(f_{pos}, t_{pos}^i) = \frac{f_{pos} \cdot t_{pos}^i}{\|f_{pos}\| \|t_{pos}^i\|}, \quad (3)$$

where  $\hat{y}$  is the predicted view class label,  $\|\cdot\|$  denotes the  $L2$  normalization. During training,  $y_i$  is the ground truth label,  $\hat{y}_i$  is the predicted label. The model is optimized using the soft cross-entropy loss as follows:

$$\mathcal{L}_{SCE} = - \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\hat{y}_{i,c}). \quad (4)$$

## 2.2 Negation Semantic-Enhanced OOD Detector

To effectively identify and reject OOD echocardiography videos, we further introduce the Negation Semantic-Enhanced (NSE) OOD Detector in our EchoView-CLIP. In NSE, we introduce another distinct visual expert  $V_{neg}$  for learning negative semantics, while sharing other parameters with the positive visual encoder  $V_{share}$ . This design prevents the negative expert from learning a distribution too similar to the positive encoder. By separating expert representations and integrating the same TML as positive encoder, our framework enhances robustness and more effectively filters out OOD samples.

Specifically, for each standard view class  $cls$ , we construct negative sentences ‘This ultrasound video does not show an  $\{cls\}$ ’ and feed them into the negative text encoder to obtain the negative text representation  $\mathcal{T}_{neg} = \{t_{neg}^1, t_{neg}^2, \dots, t_{neg}^N\}$ . Then we obtain the frame features  $h_{neg}$  from our negative visual encoder  $V_{neg}(\cdot)$ . Next, we calculate the cosine similarity between the visual features from negative expert  $h_{neg}$  and the negative text representation  $\mathcal{T}_{neg}$  to obtain a negative score for each view class, which is similar to the SVC branch. During training, we freeze the SVC branch obtained from training stage 1 (see Sec. 2.1) and optimize the  $V_{neg}$ ,  $TML_{neg}$  and  $E_{neg}$  using the soft cross-entropy loss  $\mathcal{L}_{SCE}$  and text semantic-opposite loss  $\mathcal{L}_{TSO}$  [19] as follows:

$$\mathcal{L}_{TSO} = \frac{1}{N} \left( 2 - \sum_{i=1}^N \|T_{pos(i)} - T_{neg(i)}\|_2 \right), \quad (5)$$

where  $\mathcal{L}_{TSO}$  encourages the positive and negative text features to be well-separated in the feature space. During inference, we introduce a threshold-based rejection mechanism that leverages both positive and negative similarity scores ( $Sim_{pos}^{(c)}$  and  $Sim_{neg}^{(c)}$ , respectively) for each class  $c$ . Specifically, OOD samples are rejected if the sum of class-wise scores falls below a predefined threshold:

$$\sum_{c=1}^N \left( Sim_{pos}^{(c)} \times (1 - Sim_{neg}^{(c)}) \right) < \tau, \quad (6)$$

where  $N$  is the total number of view classes, and  $\tau$  is a threshold with a default setting of 0.5 that can be adjusted as needed.

## 2.3 Quality Review for In-Distribution Videos

Some ID videos may still be of low quality due to factors such as pulmonary air interference, rib shadowing, or insufficient coupling agent, thereby cannot provide evidence to support diagnosis. To further screen out these samples to improve the data quality, we leverage the positive and negative visual semantics to assist in quality control within the quality review branch. Specifically, we introduce a new visual expert  $V_{qc}$ , which then integrates the positive and negative visual semantics through another TML module. The combined visual representations are then passed through a classifier to assess the image quality.

**Table 1.** Comparison of ours with previous methods on both standard view and OOD recognition. ‘†’ refers to the image input methods and others are video input methods.

| Model                | Standard View Recognition |              |              |              | OOD Recognition |              |              |
|----------------------|---------------------------|--------------|--------------|--------------|-----------------|--------------|--------------|
|                      | Acc.↑                     | Prec.↑       | Rec.↑        | F1↑          | AUROC↑          | FPR95↓       | Acc.↑        |
| ResNet50† [3]        | 0.882                     | 0.837        | 0.837        | 0.834        | 0.956           | 0.054        | 0.843        |
| EfficientNetV2† [17] | 0.888                     | 0.834        | 0.840        | 0.835        | 0.955           | 0.060        | 0.852        |
| ConvNeXt† [23]       | 0.896                     | 0.845        | 0.847        | 0.843        | 0.976           | 0.046        | 0.861        |
| VideoMAE v2 [20]     | 0.910                     | 0.879        | 0.893        | 0.883        | 0.981           | 0.021        | 0.897        |
| EchoPrime [18]       | 0.903                     | 0.900        | 0.903        | 0.900        | <b>0.998</b>    | 0.006        | 0.903        |
| ViFi-CLIP [14]       | 0.948                     | 0.927        | 0.930        | 0.927        | 0.986           | 0.009        | 0.945        |
| <b>Ours</b>          | <b>0.968</b>              | <b>0.958</b> | <b>0.958</b> | <b>0.957</b> | 0.993           | <b>0.002</b> | <b>0.961</b> |

**Table 2.** Ablation Study. MIL denotes original attention-based multi-instance learning module. TML means our temporal multi-instance learning module, NSE means the negation semantic-enhanced branch.

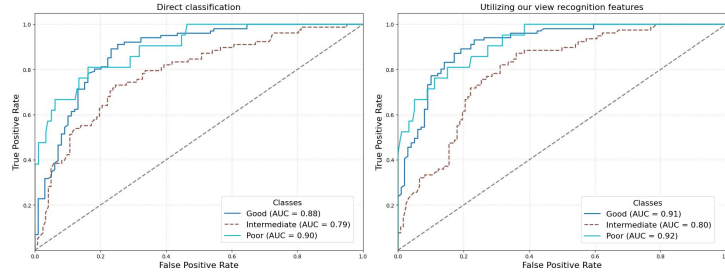
| Settings        | MIL      | TML      | NSE      | Standard View Recognition |              |              |              | OOD Recognition |              |              |
|-----------------|----------|----------|----------|---------------------------|--------------|--------------|--------------|-----------------|--------------|--------------|
|                 |          |          |          | Acc.↑                     | Prec.↑       | Rec.↑        | F1↑          | AUROC↑          | FPR95↓       | Acc.↑        |
| <b>Baseline</b> | <b>✗</b> | <b>✗</b> | <b>✗</b> | 0.948                     | 0.927        | 0.930        | 0.927        | 0.986           | 0.009        | 0.945        |
| <b>a</b>        | <b>✓</b> | <b>✗</b> | <b>✗</b> | 0.955                     | 0.948        | 0.945        | 0.946        | 0.985           | 0.007        | 0.948        |
| <b>b</b>        | <b>✓</b> | <b>✓</b> | <b>✗</b> | 0.968                     | 0.958        | 0.958        | 0.957        | 0.988           | 0.006        | 0.949        |
| <b>Ours</b>     | <b>✓</b> | <b>✓</b> | <b>✓</b> | <b>0.968</b>              | <b>0.958</b> | <b>0.958</b> | <b>0.957</b> | <b>0.993</b>    | <b>0.002</b> | <b>0.961</b> |

The output quality label is one of: good, middle, or poor. This quality review branch enhances the model’s robustness by not only identifying OOD samples but also filtering out low-quality ID samples.

### 3 Experiments

**Dataset.** We collect 20617 echocardiography videos sourcing from Guangdong Provincial People’s Hospital in China and include labels for 38 standard cardiac view classes [8], as well as the OOD category, meticulously annotated by a panel of nine experienced cardiologists. We treat videos of the 38 standard views as ID data while the rest as OOD data. This results in 15292, 4625, 200, and 200 videos for ID training and testing, OOD training and testing, respectively. To evaluate quality control performance, a subset of the ID videos was further annotated by two independent experts. Quality labels were assigned across three levels: good (standard quality excellent for diagnosis), middle (intermediate quality yet diagnostically acceptable), and poor (low quality hindering diagnosis). This quality control subset consists of 7410 training videos and 1881 testing videos.

**Implementation Details.** For image-based methods [3,17,23], we uniformly sample 16 frames from each video, assigning each frame to the corresponding video-level label. The model was trained with a batch size of 128 for 100 epochs. During testing, we averaged the predictions from the 16 frames to obtain the final video-level prediction. For video-based methods, each video is sampled to



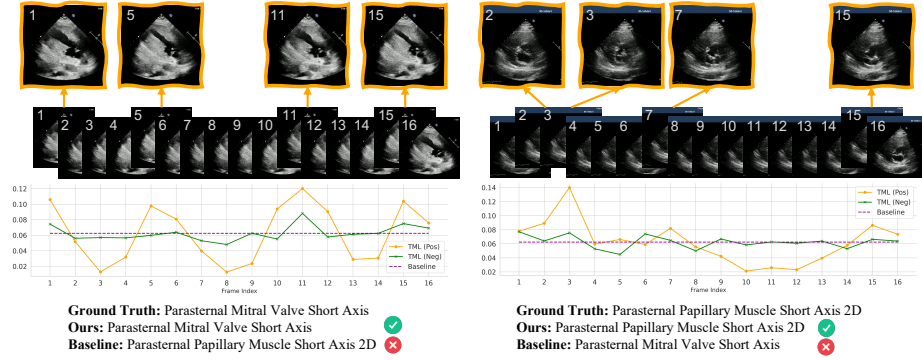
**Fig. 3.** ROC curves comparing the performance of direct classification vs. utilizing our view recognition features on CAMUS dataset.

16 frames with 1 FPS to ensure fairness. We tune EchoPrime [18] and ViFi-CLIP [14] for 30 epochs, and VideoMAEv2 [20] for 100 epoch to ensure effective adaptation. Our model was implemented using PyTorch 1.13.0 and trained on 4 NVIDIA RTX 3090 GPUs. We used the AdamW optimizer [9] with a learning rate of  $2.2 \times 10^{-5}$  and a total of 30 epochs with a batch size of 4 per GPU. For ID quality control validation, we trained 20 epoches with the same batch size.

**Standard View Recognition Performance.** We compared ours performance with a broad set of backbones and SOTA methods, including image-based methods (ResNet-50 [3], EfficientNetV2 [17], ConvNeXt [23]), and video-based methods (EchoPrime [18], ViFi-CLIP [14]). As shown in Table 1, our model achieves superior performance on standard view recognition, attaining an accuracy of 96.8% and an F1 score of 0.957, outperforming all competing methods. Notably, our model surpasses the strongest baseline, ViFi-CLIP, by a notable 2% improvement in accuracy and a 3% improvement in F1 score. For OOD detection, we follow prior work by providing additional OOD training data to models that lack inherent OOD detection capabilities. In contrast, our model is able to perform OOD detection directly by using NSE, without the need for such supplementary data. (For fair comparison, results for our method with additional OOD training data are reported in Table 2.) We evaluated performance using the following metrics: the average AUROC, FPR95 and accuracy for both ID and OOD categories in the overall test dataset. Our model achieved an exceptional AUROC of 0.993 and an FPR95 of only 0.002, demonstrating robust separation between ID and OOD samples with minimal false positives. These results meet the current diagnostic standard as confirmed by experienced clinicians, highlighting the clinical reliability and effectiveness of our model.

**Ablation Study.** We conducted ablative studies to validate the effectiveness of our proposed framework. The results are detailed in Table 2. **Setting a** demonstrates that enabling the model to focus on key frames within the input videos by using multi-instance learning module (MIL) [5] significantly improves echocardiography video recognition, boosting F1 from 0.927 to 0.946. **Setting b** shows that incorporating TML further enhances accuracy by from 0.955 to 0.968 and F1 from 0.946 to 0.957. This may because weighting video frames directly will





**Fig. 4.** Qualitative results demonstrate the superiority of our method over the baseline. For the easily confused Parasternal Papillary Muscle Short Axis 2D and Parasternal Mitral Valve Short Axis categories, our TML module effectively selects key frames to highlight distinguishing features, whereas the baseline model averages all frames, making it difficult to differentiate these closely related views.

confuse temporal information which is helpful for distinguishing between similar classes. Finally, by introducing the **NSE branch**, the model’s ability to differentiate OOD samples is improved while maintaining high accuracy for ID classification. This increases AUROC from 0.988 to 0.993 and significantly reduces FPR95, demonstrating the branch’s effectiveness.

**ID Quality Control Analysis.** Our model’s ability to identify both positive and negative semantics motivated us to further investigate its potential for assessing the quality of ID views. To this end, we trained the model on our in-house dataset to incorporate quality control capabilities. On the test dataset, the model achieved an accuracy of 82% in quality assessment. Furthermore, to address potential subjectivity in manual quality scoring on in-house data and verify the effectiveness of our model, we also evaluated the ROC curves on the public CAMUS quality classification dataset. As shown in Fig. 3, both CAMUS studies are based on our view classification model. The difference is whether view recognition features from our ID and OOD experts are used during QC expert fine-tuning. This aims to validate that accurate view classification can improve quality control. The results demonstrate that leveraging our high-performance model and its feature distribution improves the effectiveness of the ID view quality control task, particularly for distinguishing good (0.03↑) and poor (0.02↑) quality cases.

**Case study.** We present a qualitative example to illustrate the superiority of our method compared to the SOTA baseline ViFi-CLIP. As depicted in Fig. 4, we consider two visually confused categories, Parasternal Mitral Valve Short Axis and Parasternal Papillary Muscle Short Axis 2D. Our TML module effectively identifies key frames, allowing the model to focus on distinctive features that differentiate these challenging views. In contrast, the baseline model averages all frames indiscriminately, leading to difficulties in distinguishing between these closely related categories. Specifically, the baseline method misclassifies



the Parasternal Mitral Valve Short Axis as Parasternal Papillary Muscle Short Axis 2D with an 83.8% confidence score, whereas our method correctly classifies it with a 68.8% confidence score. This result highlights our model’s enhanced discriminative capability, reducing confusion between similar view classes.

## 4 Conclusion

In this paper, we have developed EchoViewCLIP, a novel framework that enhances cardiac view recognition through a negation semantic-enhanced approach and temporal-informed multi-instance learning. This framework not only improves the accuracy of fine-grained view classification across 38 standard views but also effectively rejects OOD views, thereby enhancing clinical applicability. Additionally, our model introduces an ID quality assessment branch to ensure reliable echocardiographic analysis.

**Acknowledgments.** This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N\_HKUST654/24), as well as a grant from the National Natural Science Foundation of China (Grant No. 62306254).

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

## References

1. Al-Khatib, S.M., Stevenson, W.G., Ackerman, M.J., Bryant, W.J., Callans, D.J., Curtis, A.B., Deal, B.J., Dickfeld, T., Field, M.E., Fonarow, G.C., et al.: 2017 aha/acc/hrs guideline for management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: a report of the american college of cardiology/american heart association task force on clinical practice guidelines and the heart rhythm society. *Journal of the American College of Cardiology* **72**(14), e91–e220 (2018)
2. Christensen, M., Vukadinovic, M., Yuan, N., Ouyang, D.: Vision–language foundation model for echocardiogram interpretation. *Nature Medicine* **30**(5), 1481–1488 (2024)
3. 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)
4. Heidenreich, P.A., Bozkurt, B., Aguilar, D., Allen, L.A., Byun, J.J., Colvin, M.M., Deswal, A., Drazner, M.H., Dunlay, S.M., Evers, L.R., et al.: 2022 aha/acc/hfsa guideline for the management of heart failure: a report of the american college of cardiology/american heart association joint committee on clinical practice guidelines. *Journal of the American College of Cardiology* **79**(17), e263–e421 (2022)
5. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: *International conference on machine learning*. pp. 2127–2136. PMLR (2018)

6. Khamis, H., Zurakhov, G., Azar, V., Raz, A., Friedman, Z., Adam, D.: Automatic apical view classification of echocardiograms using a discriminative learning dictionary. *Medical image analysis* **36**, 15–21 (2017)
7. Knackstedt, C., Bekkers, S.C., Schummers, G., Schreckenberg, M., Muraru, D., Badano, L.P., Franke, A., Bavishi, C., Omar, A.M.S., Sengupta, P.P.: Fully automated versus standard tracking of left ventricular ejection fraction and longitudinal strain: the fast-efs multicenter study. *Journal of the American College of Cardiology* **66**(13), 1456–1466 (2015)
8. Lang, R.M., Badano, L.P., Mor-Avi, V., Afilalo, J., Armstrong, A., Ernande, L., Flachskampf, F.A., Foster, E., Goldstein, S.A., Kuznetsova, T., et al.: Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the american society of echocardiography and the european association of cardiovascular imaging. *European Heart Journal-Cardiovascular Imaging* **16**(3), 233–271 (2015)
9. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *International Conference on Learning Representations* (2017)
10. Madani, A., Arnaout, R., Mofrad, M., Arnaout, R.: Fast and accurate view classification of echocardiograms using deep learning. *NPJ digital medicine* **1**(1), 6 (2018)
11. Narula, S., Shameer, K., Salem Omar, A.M., Dudley, J.T., Sengupta, P.P.: Machine-learning algorithms to automate morphological and functional assessments in 2d echocardiography. *Journal of the American College of Cardiology* **68**(21), 2287–2295 (2016)
12. Naser, J.A., Lee, E., Pislaru, S.V., Tsaban, G., Malins, J.G., Jackson, J.I., Anisuzzaman, D., Rostami, B., Lopez-Jimenez, F., Friedman, P.A., et al.: Artificial intelligence-based classification of echocardiographic views. *European Heart Journal-Digital Health* **5**(3), 260–269 (2024)
13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
14. Rasheed, H., Khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Fine-tuned clip models are efficient video learners. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 6545–6554 (2023)
15. Sengupta, P.P., Huang, Y.M., Bansal, M., Ashrafi, A., Fisher, M., Shameer, K., Gall, W., Dudley, J.T.: Cognitive machine-learning algorithm for cardiac imaging: a pilot study for differentiating constrictive pericarditis from restrictive cardiomyopathy. *Circulation: Cardiovascular Imaging* **9**(6), e004330 (2016)
16. Steffner, K.R., Christensen, M., Gill, G., Bowdish, M., Rhee, J., Kumaresan, A., He, B., Zou, J., Ouyang, D.: Deep learning for transesophageal echocardiography view classification. *Scientific Reports* **14**(1), 11 (2024)
17. Tan, M., Le, Q.: Efficientnetv2: Smaller models and faster training. In: *International conference on machine learning*. pp. 10096–10106. PMLR (2021)
18. 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)
19. Wang, H., Li, Y., Yao, H., Li, X.: Clipn for zero-shot ood detection: Teaching clip to say no. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 1802–1812 (2023)

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 (CVPR). pp. 14549–14560 (June 2023)
21. Wharton, G., Steeds, R., Allen, J., Phillips, H., Jones, R., Kanagala, P., Lloyd, G., Masani, N., Mathew, T., Oxborough, D., et al.: A minimum dataset for a standard adult transthoracic echocardiogram: a guideline protocol from the british society of echocardiography. *Echo Research & Practice* **2**(1), G9–G24 (2015)
22. Wilcox, J.E., Fang, J.C., Margulies, K.B., Mann, D.L.: Heart failure with recovered left ventricular ejection fraction: Jacc scientific expert panel. *Journal of the American College of Cardiology* **76**(6), 719–734 (2020)
23. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: Convnext v2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16133–16142 (2023)
24. Yang, J., Huang, T., Ding, S., Xu, X., Zhao, Q., Jiang, Y., Guo, J., Pu, B., Zheng, J., Zhang, C., et al.: Ai-enabled accurate non-invasive assessment of pulmonary hypertension progression via multi-modal echocardiography. *arXiv preprint arXiv:2505.07347* (2025)
25. 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)
26. Yang, J., Lin, Y., Pu, B., Li, X.: Bidirectional recurrence for cardiac motion tracking with gaussian process latent coding. *Advances in Neural Information Processing Systems* **37**, 34800–34823 (2024)
27. Zhang, J., Gajjala, S., Agrawal, P., Tison, G.H., Hallock, L.A., Beussink-Nelson, L., Lassen, M.H., Fan, E., Aras, M.A., Jordan, C., et al.: Fully automated echocardiogram interpretation in clinical practice: feasibility and diagnostic accuracy. *Circulation* **138**(16), 1623–1635 (2018)
28. Zheng, Z., Yang, J., Ding, X., Xu, X., Li, X.: Gl-fusion: Global-local fusion network for multi-view echocardiogram video segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 78–88. Springer (2023)