



UWL REPOSITORY

repository.uwl.ac.uk

From relative annotation to pairwise learning for assessing Echocardiographic Rotational Alignment

Alibakhshi, Alireza ORCID logoORCID: <https://orcid.org/0009-0008-7729-7320>, Fernandes, Patricia, Serej, Nasim Dadashi, Francis, Darrel P. and Zolgharni, Massoud ORCID logoORCID: <https://orcid.org/0000-0003-0904-2904> (2026) From relative annotation to pairwise learning for assessing Echocardiographic Rotational Alignment. Ultrasound in Medicine & Biology. ISSN 03015629 (In Press)

<https://doi.org/10.1016/j.ultrasmedbio.2026.09.003>

This is the Published Version of the final output.

UWL repository link: <https://repository.uwl.ac.uk/id/eprint/15541/>

Alternative formats: If you require this document in an alternative format, please contact: open.research@uwl.ac.uk

Copyright: Creative Commons: Attribution 4.0

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy: If you believe that this document breaches copyright, please contact us at open.research@uwl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Rights Retention Statement:



Contents lists available at ScienceDirect

Ultrasound in Medicine & Biology

journal homepage: www.elsevier.com/locate/ultrasmedbio

Original Contribution

From Relative Annotation to Pairwise Learning for Assessing Echocardiographic Rotational Alignment

Alireza Alibakhshi^{a,c,*}, Patricia Fernandes^a, Nasim Dadashi Serej^{a,b}, Darrel P. Francis^c, Massoud Zolgharni^{a,b,c}^a Translational Healthcare Research Centre, University of West London, London, UK^b School of Computing and Engineering, University of West London, London, UK^c National Heart and Lung Institute, Imperial College London, London, UK

ARTICLE INFO

Keywords:

Echocardiography
Qualitative image quality assessment
Rotational misalignment
Pairwise learning
Relative annotation
Expert consensus

ABSTRACT

Objective: To evaluate whether preserving relative expert-derived structure during model training improves the automated assessment of rotational misalignment in apical four-chamber echocardiographic images.

Methods: A multi-expert dataset of apical four-chamber images was annotated using multiple-image ranking to derive consensus rotational orderings along a clockwise-to-anti-clockwise continuum. Conventional pointwise regression models were compared with shared-weight pairwise models that received two images and predicted the signed difference between their consensus rotational scores. Performance was evaluated across three independent expert-ranked test sets using pairwise agreement and Spearman correlation. Robustness was further assessed using leave-one-expert-out and cycle-consistency analyses, with saliency-based visualization used to examine the image regions influencing predictions.

Results: Pairwise learning increased the mean pairwise agreement from 74.86% to 84.29% and the mean Spearman correlation from 67.74% to 85.43% compared with pointwise regression. The best pairwise model achieved a pairwise agreement of up to 87.81% and a Spearman correlation of up to 91.23%. The performance remained stable when individual annotators were excluded, and a cycle-consistency analysis demonstrated few cyclic preferences on independent test data. Visual explanation maps indicated model attention to clinically relevant cardiac anatomy.

Conclusion: Training directly on relative expert-derived structure improved agreement with multi-expert consensus compared with conventional pointwise regression. Pairwise learning therefore provides a robust framework for assessing qualitative, continuous echocardiographic image-quality attributes such as rotational alignment.

Introduction

Echocardiography is a first-line imaging modality for assessing cardiac structure and function, but its diagnostic value depends strongly on the quality and standardization of image acquisition. The apical four-chamber (A4C) view is particularly important because it underpins visual interpretation and many downstream measurements, including chamber dimensions, ventricular volumes, ejection fraction and longitudinal strain. However, the acquisition of an optimal A4C view remains operator dependent. Patient habitus, acoustic windows, rib shadowing, breathing, machine settings and probe handling can all alter the apparent anatomy and reduce the reliability of visual and quantitative assessments.

Image quality in echocardiography is not a single property. It includes anatomical visibility, technical signal quality, temporal stability, foreshortening, centering and acquisition correctness. Rotational

misalignment is one component of acquisition correctness. In the A4C view, rotation occurs when the transducer is twisted clockwise or anti-clockwise around its imaging axis, causing the imaging plane to drift away from the intended four-chamber plane toward neighboring apical configurations. An A4C image may therefore be technically clear and correctly classified as A4C, yet still be displaced along a clockwise-anti-clockwise alignment axis that changes the apparent relationship between the ventricles, atria and atrioventricular valves. Despite its clinical relevance, rotational misalignment has received less attention than broader image-quality scoring, view classification or foreshortening assessment. Detecting such misalignment could support a more standardized acquisition, provide feedback during training and help to identify images that may be less reliable for downstream automated measurement. The present study focused specifically on apparent clockwise-anti-clockwise rotational alignment within images already identified as A4C views and does not independently assess other probe

* Corresponding author. Translational Healthcare Research Centre, University of West London, St Mary's Rd, Ealing, London W5 5RF, UK.

E-mail address: alireza.alibakhshi@uwl.ac.uk (A. Alibakhshi).

<https://doi.org/10.1016/j.ultrasmedbio.2026.09.003>

Received 19 May 2026; Revised 28 August 2026; Accepted 2 September 2026

0301-5629/© 2026 The Author(s). Published by Elsevier Inc. on behalf of World Federation for Ultrasound in Medicine and Biology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

manipulations such as tilt, translation, out-of-plane displacement or foreshortening.

A correctly aligned A4C view depends on the coherent presentation of several anatomical landmarks, including the interventricular septum, the left and right ventricular cavities, the atrioventricular valve plane and the atria. Rotational misalignment can alter these relationships without necessarily degrading the overall image brightness or contrast. This quality makes rotation different from purely technical image-quality defects: it is primarily a geometric and anatomical alignment problem rather than simply a signal-quality problem and may therefore affect downstream interpretation even when image brightness and contrast appear to be acceptable.

The broader problem of the A4C orientation has previously been recognized in our work on on-axis imaging. Dhutia et al. [1] developed an automated speckle-tracking algorithm to detect medial–lateral probe translation in A4C echocardiography and showed that the algorithm-derived orientation assessment agreed with specialist consensus at a level comparable with individual specialists. Their work established three principles that are also relevant here: first, that correct A4C acquisition is partly subjective; second, that expert consensus is a pragmatic reference standard when no independent quantitative gold standard exists; and third, that objective image-orientation metrics may support training, quality control and acquisition guidance. The present study extends this line of work toward a complementary within-view alignment problem: the clockwise–anti-clockwise rotational misalignment of the A4C view. Representative examples of this rotational alignment continuum are shown in Figure 1.

A central difficulty is that rotational misalignment is not naturally judged as an isolated absolute score. Experts often judge whether one image is more clockwise or more anti-clockwise than another, or whether an image is closer to a familiar reference appearance. This makes the task inherently comparative. Conventional artificial intelligence (AI) pipelines, however, usually collapse expert judgment into a single label or continuous score for each image, then train a model to predict that pointwise target. This process creates a mismatch between the way experts provide judgment and the way the model is trained.

Relative annotation has been proposed as one way to reduce this mismatch. Zaman et al. [2] introduced a multiple-image ranking strategy in cardiac magnetic resonance imaging, showing that expert time could be used more efficiently when clinicians ranked images relative to one another rather than labeling each image in isolation. Our previous work translated this idea to echocardiographic A4C rotational misalignment: experts ranked A4C images from anti-clockwise to clockwise and the resulting consensus was used to train a regression model [3]. That study demonstrated the feasibility of rotational misalignment assessment in echocardiography, but the learning stage remained pointwise: the model still received a single image and learned to predict an absolute score.

This problem is not unique to A4C rotation. Many echocardiographic quality attributes are qualitative, continuous and difficult to score

absolutely, including endocardial border visibility, anatomical coverage, centering, foreshortening and task readiness for downstream measurement. In these settings, experts may be more consistent when making comparative judgments than when assigning isolated numerical scores. A4C rotational misalignment therefore provides a focused and clinically meaningful case study for testing whether pairwise learning can better preserve expert judgment in echocardiographic quality assessment.

In the present study, we go beyond relative annotation and reformulate the learning problem itself as a pairwise comparison. We demonstrate this using A4C rotational misalignment as a representative qualitative acquisition-alignment task. Instead of using relative annotations only to create a global score and then training a conventional regressor, we train a shared-weight pairwise model to compare two A4C images and predict the signed difference between their consensus-derived rotational scores. The predicted pairwise relationships are then aggregated into a global ranking. This design preserves relative structure during model learning while using the aggregated expert consensus as a denoised supervision signal. More broadly, it provides a methodological framework that may generalize to other qualitative echocardiographic quality attributes where expert judgment is naturally comparative.

Related work

Deep learning has been widely applied to echocardiography for view classification, segmentation, automated measurement and image-quality assessment. Existing quality control approaches have considered global image quality, anatomical visibility, gain and contrast, foreshortening and suitability for downstream analysis [4–6]. However, echocardiographic image quality is multidimensional: an image may have adequate signal quality while remaining anatomically misaligned. For the A4C view, acquisition geometry is therefore important in addition to structure visibility.

Previous work on standard-view acquisition has incorporated probe manipulation, including rotation and tilt, into real-time guidance and scan-assistance systems [7,8]. These approaches demonstrate the importance of acquisition geometry, but generally address navigation toward standard views or combined quality feedback rather than continuous clockwise–anti-clockwise rotational alignment within an established A4C view. Rotational misalignment is particularly suited to comparative assessment because the visual differences may be subtle and experts may find it easier to judge the relative orientation of two images than to assign an absolute numerical score.

Pairwise comparison and rank learning have been used in general image-quality assessment when absolute ratings are difficult to define reliably [9–11]. Similar advantages have been reported in medical imaging, where pairwise comparison improved observer agreement compared with conventional rating scales [12]. Multiple-image ranking extends this principle by allowing experts to order several examples simultaneously and thereby generate multiple relative constraints

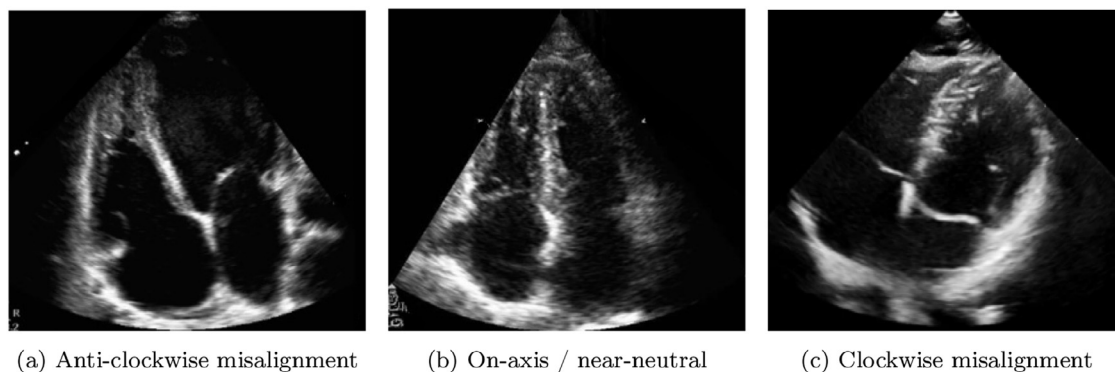


Figure 1. Representative examples illustrating the rotational alignment continuum in apical four-chamber echocardiography, ranging from (a) anti-clockwise rotational misalignment, through (b) near-neutral alignment, to (c) clockwise rotational misalignment.

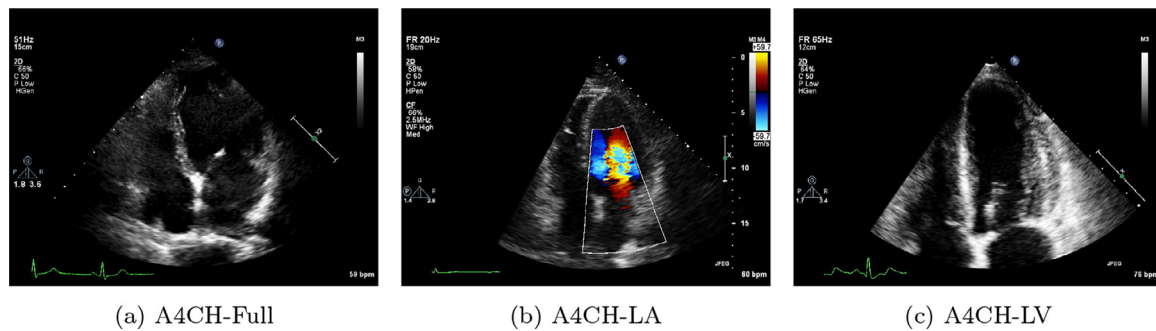


Figure 2. Representative examples of the three apical four-chamber (A4C)–related subview categories retained in this study. (a) *A4CH-Full* shows the full depth of all four chambers, (b) *A4CH-LA* shows the full depth of the left atrium but not the full depth of the left ventricle and (c) *A4CH-LV* shows the full depth of the left ventricle but not the full depth of the left atrium.

efficiently. This approach has previously been applied in cardiac magnetic resonance imaging [2].

In our previous work, multiple-image ranking was used to derive a consensus rotational scale for A4C echocardiographic images, but the resulting model was trained using conventional pointwise regression [3]. The present study extends the relative formulation from annotation to model learning by training a shared-weight pairwise network on consensus-derived rotational differences. We therefore directly compare pointwise regression with pairwise learning under the same expert-derived supervision and evaluation framework, testing whether preserving the comparative structure during learning improves agreement with expert rotational judgment.

Key contributions

This study makes the following contributions.

- **Pairwise learning for A4C rotational alignment assessment:** We extend pairwise and rank-learning ideas to echocardiography by introducing a shared-weight model that compares two A4C images and predicts their relative clockwise–anti-clockwise displacement using consensus-derived pairwise targets.
- **Direct comparison with pointwise regression:** We compare pairwise learning with conventional pointwise regression across baseline and echocardiography–pre-trained backbones using the same expert-derived consensus data.
- **UNITY-ROT benchmark:** We introduce *UNITY-ROT*, a publicly available benchmark for A4C rotational misalignment, comprising expert-ranked A4C images and independent multi-expert evaluation sets for the reproducible assessment of clockwise–anti-clockwise alignment models.
- **Robust expert-consensus validation:** We evaluate performance on three independent 30-image expert-ranked evaluation sets and perform leave-one-expert-out analyses to assess sensitivity to individual annotators.

Methods

This section describes the data curation, expert annotation, model architectures, training protocol and evaluation strategy. The key methodological distinction is between the pointwise regression baseline, which predicts a continuous rotational score for a single image and the proposed pairwise model, which learns the relative rotational ordering of two images.

Patient dataset

Echocardiographic studies were obtained from the Imperial College Healthcare NHS Trust echocardiogram database under Health Research

Authority approval for the anonymized export of imaging data originally acquired during routine clinical care. Studies were acquired by experienced sonographers using standard clinical protocols and ultrasound systems from multiple manufacturers. All patient-identifiable metadata were removed before export.

The source database contains real-world transthoracic echocardiographic studies acquired across a heterogeneous clinical population and a range of scanner systems. This heterogeneity was important for the present study, as rotational misalignment is influenced by patient anatomy, acoustic windows, probe position and acquisition practice. The aim was therefore to derive a dataset that reflected the visual variability encountered in routine echocardiography practice rather than a narrowly curated set of ideal images.

Data preparation

Candidate apical-view videos were first identified using a video view classification framework [13]. Figure 2 shows representative examples of the three A4C-related subviews. Three A4C-related categories were retained for this study:

- **A4CH-Full:** 4080 videos;
- **A4CH Left Atrium (A4CH-LA):** 3079 videos; and
- **A4CH Left Ventricle (A4CH-LV):** 3744 videos.

Each video contained one to three cardiac cycles. Representative frames were selected using a cardiac phase detection model [14]. For each video, the end-diastolic and end-systolic frames were identified and two additional intermediate frames were sampled halfway between these phases. This strategy produced four representative frames per video, capturing cardiac morphology at different points across the cardiac cycle.

To prioritize informative examples, reduce redundant annotation and ensure that experts reviewed a diverse set of cases, a feature-based sample-selection step was applied. Features were extracted from the four representative frames using a convolutional neural network feature extractor and concatenated into a single video-level vector. Principal component analysis followed by clustering was then used to identify representative videos spanning the observed range of appearance. The most diverse and representative cases were prioritized for expert annotation.

Ground truth definition and ranking protocol

The annotation protocol was designed to capture rotational misalignment as a relative judgment rather than an isolated absolute score. Following the diversity-based selection step, the top 200 A4CH-Full videos and the top 100 videos from each of the A4CH-LA and A4CH-LV categories were selected, yielding 400 videos for annotation. Four

Algorithm 1

Consensus-based Rotational Misalignment Annotation.

Input : Image set $\mathcal{I} = \{1, \dots, N\}$; batch size $B = 6$; minimum views $V_{\min} = 3$.

Output: Consensus ranking π , ratings $\{r_i\}$, views $\{v_i\}$, volatilities $\{\phi_i\}$.

Randomly shuffle $\mathcal{I}$ to obtain initial ranking indices $\{\text{rank}_i\}$.

```

foreach  $i \in \mathcal{I}$  do
   $r_i \leftarrow 1500$ ;  $v_i \leftarrow 0$ ;  $\phi_i \leftarrow 0.5$ 
while  $\exists i \in \mathcal{I} : v_i < V_{\min}$  do
  foreach  $i \in \mathcal{I}$  do
     $P_i \leftarrow \mathbf{1}[v_i \geq V_{\min}] + \lambda_1 v_i - \lambda_2 \phi_i + \varepsilon_i$ 
   $\mathcal{B} \leftarrow \arg \min_{\text{items}} \{P_i \mid i \in \mathcal{I}\}$ 
   $\text{ORDER} \leftarrow \text{PresentDragDrop}(\mathcal{B})$ 
  foreach  $i \in \text{ORDER}$  do
     $v_i \leftarrow v_i + 1$ 
  for  $p \leftarrow 1$  to  $B$  do
    for  $q \leftarrow p + 1$  to  $B$  do
       $a \leftarrow \text{ORDER}[p]$ ;  $b \leftarrow \text{ORDER}[q]$ 
       $\hat{p}_{b>a} \leftarrow \sigma\left(\frac{r_b - r_a}{S}\right)$ 
       $r_b \leftarrow r_b + K(1 - \hat{p}_{b>a})$ ;  $r_a \leftarrow r_a - K(1 - \hat{p}_{b>a})$ 
   $\pi \leftarrow \text{argsort}_{i \in \mathcal{I}} \text{ of } r_i \text{ descending}$ 
  foreach  $i \in \mathcal{B}$  do
     $\phi_i \leftarrow (1 - \alpha)\phi_i + \alpha \text{MapTo}_{[0,1]}(\tau)$ 
return  $\pi, \{r_i\}, \{v_i\}, \{\phi_i\}$ 

```

representative frames were extracted from each video, producing 1600 images for expert annotation.

These images were annotated by a pool of 25 clinical experts using our custom-built, web-based annotation platform: <https://unityimaging.net/>. Experts were presented with randomly shuffled batches of six images and asked to reorder them from most anti-clockwise to most clockwise according to perceived A4C rotational alignment. This drag-and-drop multiple-image ranking strategy enabled experts to compare images directly rather than score them individually. This methodology builds on prior work showing that relative labeling can generate robust consensus labels from limited expert time [2].

The Unity platform dynamically selected images for annotation using four internal variables that were not visible to annotators.

- **Ranking**: the current position of each image within the dataset, initialized randomly.
- **Rating**: a continuous algorithmic score initialized at 1500, the neutral midpoint of the 0–3000 scale and updated after each annotation batch. Lower values indicated greater anti-clockwise rotation, while higher values indicated greater clockwise rotation.
- **View count**: the number of times each image had been annotated.
- **Volatility**: a measure of annotation instability, ranging from 0 (stable/high agreement) to 1 (high disagreement).

These parameters were updated in real time after each batch submission. Images with low view counts or high volatility were prioritized until all images had appeared in at least three annotation batches.

Experts independently annotated overlapping subsets and their responses were aggregated to produce a consensus-based global ranking and continuous rating of rotational alignment.

Algorithm 1 summarizes the annotation and consensus-generation procedure and Figure 3 illustrates the data curation and annotation pipeline.

Development and evaluation datasets

The annotated development dataset comprised 1600 images extracted from 400 videos. These images were used for model development, including training, fine-tuning, internal validation and internal testing. The dataset was divided into three non-overlapping subsets: 1300 images for training, 200 for internal validation and 100 for internal testing. To prevent data leakage, images from the same echocardiographic study were restricted to a single subset.

For the final evaluation against the expert consensus, three additional expert-ranked evaluation sets were curated. Each set contained 30 images randomly selected from the 4080 A4CH-Full videos. These images were entirely excluded from the development dataset.

Each 30-image evaluation set was presented in full to 10 clinical experts, who independently reordered the images from most anti-clockwise to most clockwise according to perceived rotational alignment. This produced 10 independent expert rankings per set, which were aggregated to form reference consensus rankings for evaluating model performance. The use of independent multi-expert rankings provided a clinically meaningful assessment of agreement between model predictions and human judgment.

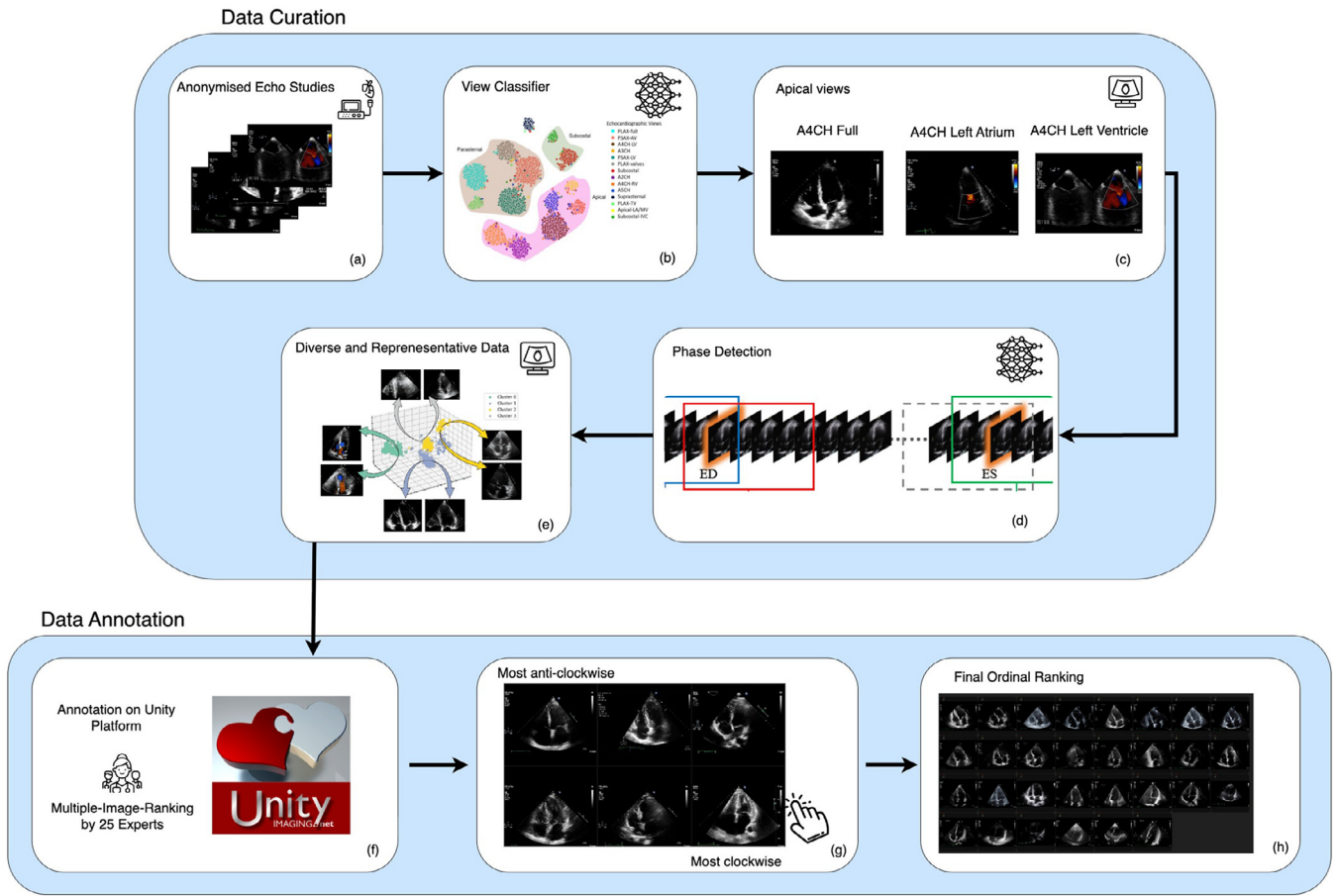


Figure 3. Overview of the data curation and annotation pipeline. (a) A large random sample of echocardiographic studies was extracted from the NHS echocardiography database and anonymized. (b, c) Relevant apical views were automatically identified using a convolutional neural network–based view classifier [15]. (d) End-diastolic (ED) and end-systolic (ES) frames were detected using our validated cardiac phase detection model [16]. (e) Features extracted from frames at defined points in the cardiac cycle were used to represent each patient's video and clustered, enabling the selection of diverse and representative cases for expert annotations [14]. (f) A multiple-image ranking strategy was implemented in the Unity labeling platform to annotate rotational misalignment [2]. (g) Shuffled images were presented to annotators in batches of six and annotators reordered them by dragging image tiles from "most anti-clockwise" (top left) to "most clockwise" (bottom right). (h) The final output was an ordinal ranking of images reflecting the degree of rotational misalignment. A4CH-Full shows the full depth of all four chambers. A4CH Left Atrium shows the full depth of the left atrium but not the full depth of the left ventricle. A4CH Left Ventricle shows the full depth of the left ventricle but not the full depth of the left atrium.

Model backbones

We evaluated each learning formulation using both baseline and echocardiography pre-trained feature extractors. This allowed us to separate the effect of the learning objective from the effect of representation quality.

As a baseline convolutional feature extractor, we used a residual network architecture. Residual networks use skip connections to support the training of deeper models by learning residual mappings rather than entirely new transformations at each layer. In this study, the ResNet-based model provided a non–echocardiography-specific baseline for assessing whether pairwise learning could improve performance even without specialized pre-training.

We also evaluated Xception, which uses depthwise separable convolutions to separate spatial filtering from cross-channel mixing. This design reduces computational costs while preserving a strong feature-extraction capacity, making it suitable for medical imaging tasks where robust representations are needed but labeled data may be limited.

To assess the value of echocardiography-specific pre-training, we compared baseline weights with backbones pre-trained on echocardiographic view-classification tasks. Specifically, we evaluated

models pre-trained on 14-class video-level view classifications [13] and 47-class fine-grained view classifications [15]. Such pre-training encourages the model to represent cardiac anatomy and view-specific morphology before adaptation to rotational misalignment. This is relevant, because cues for A4C rotational alignment depend on the relative appearance of the LV, right ventricle, atria and valve plane.

Learning formulations

We compared two learning formulations: conventional pointwise regression and the proposed pairwise comparison model. Both formulations used the same development data and were evaluated under the same expert-consensus framework.

Pointwise regression

In the pointwise regression baseline, each image was treated independently. Given an echocardiographic image x , a backbone feature extractor f_{bb} produced an embedding that was passed through a regression head to predict a continuous rotational score,

$$\hat{r} = f_{reg}(x). \quad (1)$$

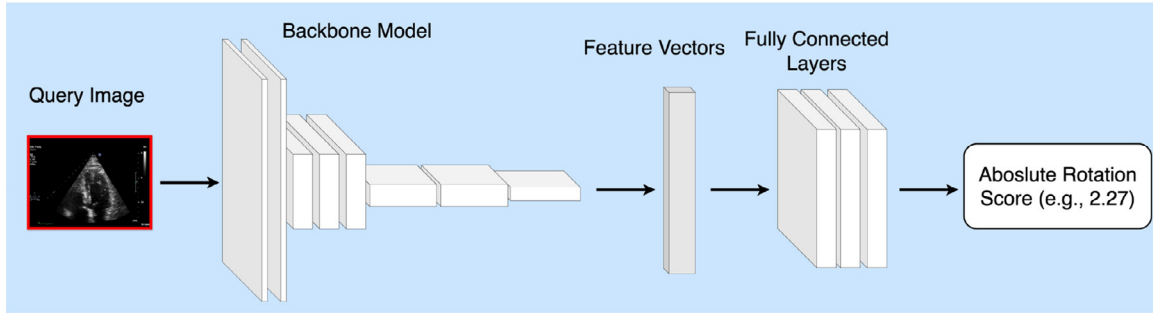


Figure 4. Pointwise regression model. A single apical four-chamber (A4C) image is passed through a backbone feature extractor and regression head to predict a continuous rotational score. Global ordering is obtained by sorting the predicted scores.

The output was constrained to the normalized expert-derived rating range. At inference, images were globally ranked by sorting their predicted scores. This formulation therefore requires the model to estimate an image-level absolute score, despite the underlying expert judgment being primarily comparative. The pointwise regression architecture is illustrated in Figure 4.

Pairwise comparison learning

The pairwise formulation was designed to preserve the relative structure of expert rotational judgment during model learning. Expert annotations were first aggregated into a consensus rotational scale and pairwise supervision was subsequently generated from differences between the corresponding consensus scores.

Two images, x_i and x_j , were processed by identical shared-weight feature extractors. Their latent representations were compared using a signed difference vector and a lightweight regression head predicted their relative rotational displacement:

$$\hat{\Delta}_{ij} = g_{\phi} [f_{\theta}(x_i) - f_{\theta}(x_j)]. \quad (2)$$

The corresponding supervision target was defined as

$$\Delta_{ij} = y_j - y_i, \quad (3)$$

where lower consensus scores corresponded with greater anti-clockwise rotation and higher scores with greater clockwise rotation. Thus, $\Delta_{ij} > 0$ indicates that image i is more anti-clockwise than image j , whereas $\Delta_{ij} < 0$ indicates the reverse ordering.

The model was trained end-to-end by minimizing the regression loss between the predicted and expert-consensus pairwise differences. At inference, the sign of each predicted difference defined the preferred ordering for that image pair. Pairwise preferences were then aggregated using an Elo-style procedure to produce a global ranking, with higher final Elo ratings corresponding with more anti-clockwise images. For efficient inference, each image embedding can be computed once and reused across all pairwise comparisons.

The pairwise architecture does not explicitly enforce antisymmetry or transitivity; the consistency of the resulting pairwise predictions was therefore assessed empirically using a cycle-consistency analysis. Extended mathematical details and the Elo aggregation algorithm are provided in the Supplementary methods.

Training and inference

The pairwise architecture, as shown in Figure 5, does not explicitly enforce antisymmetry or transitivity; the consistency of the resulting pairwise predictions was therefore assessed empirically using cycle-consistency analysis. Extended mathematical details and the Elo aggregation algorithm are provided in the Supplementary Methods.

All models were implemented in TensorFlow 2.0 and trained on an NVIDIA GeForce RTX 4090 GPU. Frames were zero-padded and resized

to 256×256 pixels. The annotated development dataset comprised 1600 images and was divided into three non-overlapping subsets: 1300 images for training, 200 for internal validation and 100 for internal testing. Images from the same echocardiographic study were kept within a single subset to reduce the risk of data leakage.

For the pointwise regression baseline, each image was presented independently and the model was trained to predict the expert-derived continuous rotational rating. Models were trained for up to 20 epochs with a batch size of 32, using validation loss and early stopping to reduce overfitting.

For the pairwise model, the data loader generated image pairs from the training subset only. For a pair of frames (x_i, x_j) with expert-derived consensus ratings (y_i, y_j) , the pairwise target was defined as the signed rating difference:

$$\Delta_{ij} = y_j - y_i. \quad (4)$$

Because lower rating values correspond with a more anti-clockwise rotation and higher values correspond with a more clockwise rotation, this convention means that $\Delta_{ij} > 0$ indicates that image i is more anti-clockwise than image j , whereas $\Delta_{ij} < 0$ indicates that image i is more clockwise than image j . Pairwise models were trained for up to 20 epochs with a batch size of 8 and early stopping. The raw expert rankings were therefore used to construct a consensus scale first and pairwise training targets were generated from that consensus scale.

At inference, the regression model produced one score per image and global rankings were obtained by sorting these scores. For the pairwise model, all unordered image pairs in the evaluation set were passed through the shared-weight architecture to obtain predicted signed differences. The sign of each prediction was used as a pairwise preference and these preferences were aggregated using the Elo-style rating procedure in Algorithm 2 in the Supplementary materials. Thus, regression inference produced a ranking by sorting independently predicted image-level scores, whereas pairwise inference produced a ranking by aggregating the predicted relative preferences across all of the image pairs.

Evaluation metrics

Because pointwise regression and pairwise comparison produce different native outputs, a common evaluation framework was used. Evaluation was performed on the three independent expert-ranked 30-image sets described in the “Development and Evaluation Datasets” Section.

Two primary metrics were used. First, pairwise agreement measured the proportion of image pairs for which the model predicted the same relative ordering as the expert consensus. Let

$$\Delta_{ij} := y_j - y_i, \quad \mathcal{S}^{\star} := \{(i, j) : i < j, \Delta_{ij} \neq 0\}. \quad (5)$$

For the pairwise model, agreement was computed directly from the predicted signed difference $\hat{\Delta}_{ij}$:

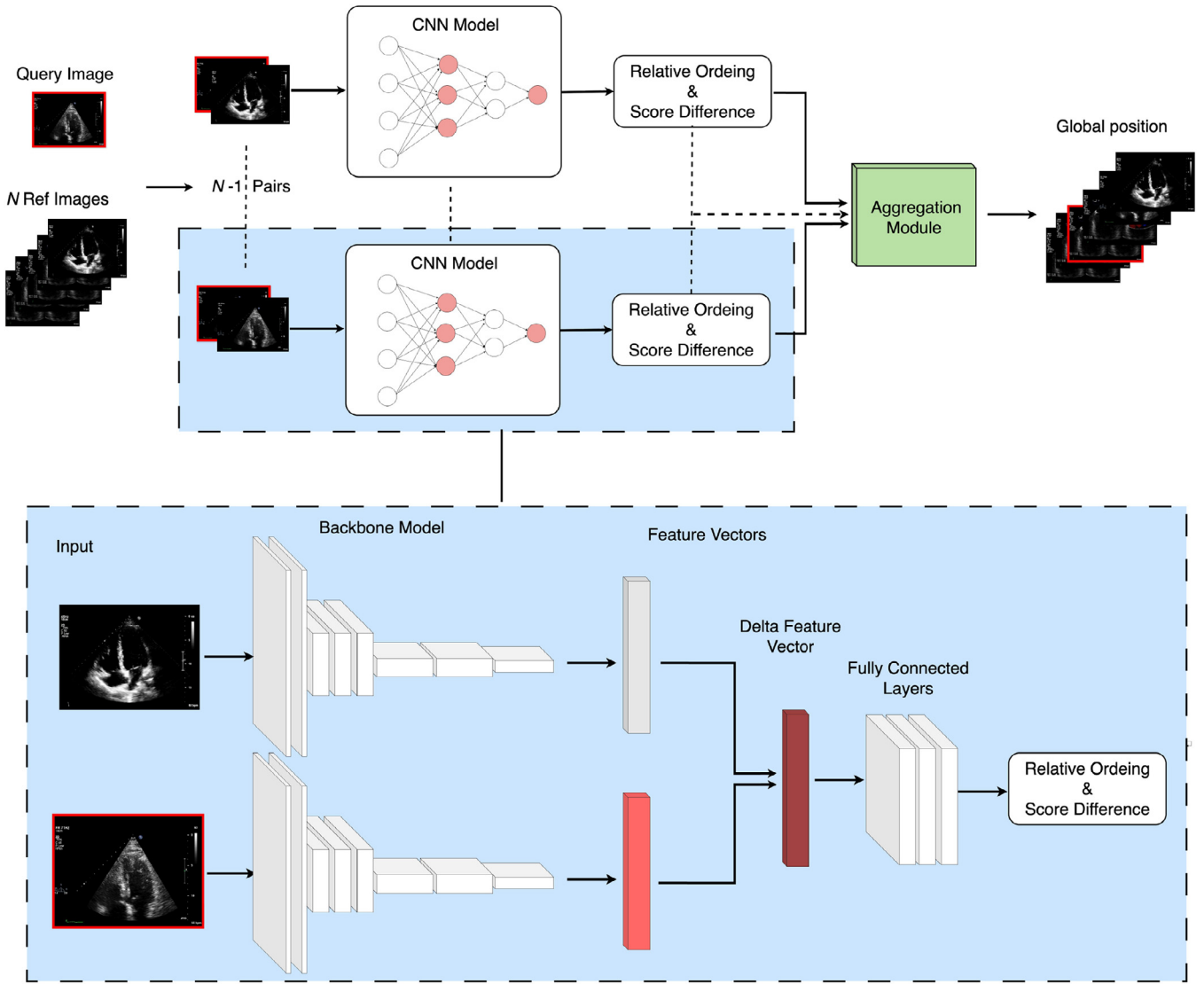


Figure 5. Pairwise comparison model. Two apical four-chamber (A4C) images are passed through shared-weight feature extractors and the signed difference between their embeddings is used to predict relative clockwise–anti-clockwise displacement. Pairwise preferences are subsequently aggregated using an Elo-style procedure to obtain the global ranking. Training targets are derived from expert consensus rotational scores. CNN, convolutional neural network.

$$\text{Accuracy}_{\text{pair}} = \frac{1}{|\mathcal{P}^{\star}|} \sum_{(i,j) \in \mathcal{P}^{\star}} \mathbb{1} \left[\left(\Delta_{ij} > 0 \wedge \hat{\Delta}_{ij} > 0 \right) \vee \left(\Delta_{ij} < 0 \wedge \hat{\Delta}_{ij} < 0 \right) \right]. \quad (6)$$

For regression models, pairwise ordering was obtained by subtracting the two predicted image-level scores using the same sign convention:

$$\hat{\Delta}_{ij}^{\text{reg}} = \hat{y}_j - \hat{y}_i. \quad (7)$$

Regression pairwise agreement was then defined as

$$\text{Accuracy}_{\text{reg}} = \frac{1}{|\mathcal{P}^{\star}|} \sum_{(i,j) \in \mathcal{P}^{\star}} \mathbb{1} \left[\left(\Delta_{ij} > 0 \wedge \hat{\Delta}_{ij}^{\text{reg}} > 0 \right) \vee \left(\Delta_{ij} < 0 \wedge \hat{\Delta}_{ij}^{\text{reg}} < 0 \right) \right]. \quad (8)$$

Second, a Spearman correlation was used to assess agreement between the model-derived global ranking and the expert consensus ranking. The pairwise model rankings were obtained by an Elo-style aggregation of predicted comparisons, whereas the regression rankings were obtained by sorting predicted image-level scores. In both cases, rankings were oriented consistently with the expert consensus, from most anti-clockwise to most clockwise.

Results and discussion

Table 1 reports the pairwise agreement across the three independent expert-ranked evaluation sets. For each 30-image set, performance was defined as the proportion of correctly predicted relative orderings across all possible image pairs. Across all three backbone settings, the pairwise formulation outperformed the corresponding pointwise regression formulation.

The strongest individual result was obtained by the pairwise baseline model on set 2 (87.81%), followed by the pairwise model with 14-class pre-training on set 1 (87.35%). Averaged across the three sets, the pairwise models achieved mean agreement values between 83.36% and 84.29%, whereas the baseline regression model achieved 74.86%. These results suggest that the learning consensus–derived relative differences between images is better aligned with the rotational misalignment task than learning only an absolute image-level score.

Table 2 shows the correlation between the model-derived rankings and the multi-expert consensus ranking. The pairwise models again achieved the highest correlations across the evaluation sets. The best individual result was obtained by the pairwise model with 14-class pre-training on set 1 (91.23%), while the pairwise model with 47-class pre-training showed the most consistent performance across the three sets (88.65%, 89.32% and 86.87%). In contrast, regression without

Table 1

Pairwise agreement of different approaches for predicting rotational misalignment

Model	Set 1	Set 2	Set 3	Mean
Regression – baseline	72.87	74.71	77.01	74.86
Regression – pre-trained-14	85.28	81.83	77.93	81.68
Regression – pre-trained-47	83.67	82.06	79.31	81.68
Pairwise – baseline	85.28	87.81^a	79.77	84.29^a
Pairwise – pre-trained-14	87.35^a	81.92	80.91	83.39
Pairwise – pre-trained-47	85.05	83.67	81.37^a	83.36

For each 30-image evaluation set, agreement was defined as the proportion of correctly predicted relative orders across all possible image pairs ($\frac{30 \times 29}{2} = 435$) compared with the multi-expert consensus ranking.

^a Bold values indicate the best-performing result for each set. (column)

Table 2

Spearman correlation between model-derived rankings and multi-expert consensus rankings

Model	Set 1	Set 2	Set 3	Mean
Regression – baseline	63.95	65.78	73.48	67.74
Regression – pre-trained-14	88.07	86.29	77.97	84.11
Regression – pre-trained-47	85.13	83.33	78.28	82.25
Pairwise – baseline	86.20	90.34^a	79.75	85.43
Pairwise – pre-trained-14	91.23^a	86.42	84.07	87.24
Pairwise – pre-trained-47	88.65	89.32	86.87^a	88.28^a

For each 30-image evaluation set, performance was measured as the correlation between the model-predicted global ranking and the corresponding expert consensus ranking.

^a Bold values indicate the best-performing result for each set. (column)

echocardiography-specific pre-training performed poorly, particularly on sets 1 and 2. Pre-training substantially improved regression, but did not remove the overall advantage of the pairwise formulation.

The results also clarify the role of pre-training. Regression models benefited strongly from echocardiography-specific pre-training, suggesting that the pointwise prediction of an absolute rotational score depends heavily on a well-structured anatomical representation. Pairwise learning appeared to be less dependent on pre-training: even the baseline pairwise model remained competitive with the pre-trained pairwise variants because the supervision signal was defined by consensus-derived relative differences between the images. This finding suggests that, for qualitative alignment attributes where expert judgment is inherently relative, the formulation of the learning problem may be as important as the feature representation itself. Overall, the strongest mean correlation was achieved by the pairwise model with 47-class pre-training, indicating that pre-training can still provide additional benefit when combined with a pairwise objective. These findings support the central hypothesis of the study: when expert judgment is comparative, preserving the relative structure during learning provides a more natural training signal than collapsing the task into pointwise regression.

Inference computational performance

Table 3 shows the computational performance evaluated for the matched pre-trained-47 models. The pairwise model improved the mean pairwise agreement from 81.68% to 83.36% and the mean Spearman correlation from 0.82 to 0.88. Based on Algorithm 2 in the Supplementary materials for efficient inference, each image embedding was computed once and cached for re-use across all pairwise comparisons, avoiding repeated backbone evaluation. With this implementation, ranking 30 images required 14.71 ± 0.13 ms for pairwise learning vs. 13.12 ± 0.08 ms for regression, despite the larger pairwise model (33.71 million vs. 24.07 million parameters). Detailed computational results for all backbones are provided in the Supplementary materials.

Table 3

Performance and computational characteristics of the matched pre-trained-47 regression and pairwise models

Metric	Regression	Pairwise
Mean pairwise agreement	81.68%	83.36%
Mean Spearman correlation	0.82	0.88
Total parameters	24.07 M	33.71 M
30-Image ranking time	13.12 ± 0.08 ms	14.71 ± 0.13 ms

Ranking times represent the model-compute time for 30 images and are reported as mean $\pm$ standard deviation over 30 repetitions.

Human expert benchmark

To contextualize model performance, we compared each individual expert ranking with the multi-expert consensus ranking for the three independent evaluation sets. Across all 30 expert observations, individual experts achieved a mean Spearman correlation of 0.927 ± 0.053 with the consensus ranking, with values ranging from 0.743 to 0.981. The corresponding mean Kendall's tau was 0.789 ± 0.078 and the mean pairwise agreement was 0.895 ± 0.039 , with the pairwise agreement ranging from 0.777 to 0.947.

The pairwise models performed within the observed range of individual expert agreement with the consensus across all three metrics, although below the average individual expert. The strongest pairwise model achieved a pooled Spearman correlation of 0.876 and a pairwise agreement of 0.838. These results indicate that the proposed pairwise formulation captures a clinically meaningful component of expert rotational judgment, while leaving room for improvement before average expert-level consensus agreement is reached.

Figure 6 shows the set-wise relationship between the model rank correlation and individual expert variability. The best pairwise models fall within the observed range of the individual expert agreement for each evaluation set, although generally below the expert mean. Because each expert contributed to the consensus ranking, this benchmark should be interpreted as a descriptive measure of expert–consensus agreement rather than an independent leave-one-expert-out human ceiling.

Robustness to individual expert exclusion

Table 4 reports the leave-one-expert-out robustness analysis for set 1. For each run, the reference consensus ranking was recalculated after excluding one expert and model correlations were recomputed against the remaining nine-expert consensus. The pairwise models remained high and stable across all exclusions. The strongest mean correlation was achieved by the pairwise model with 14-class pre-training (91.63), while both pre-trained pairwise models showed the lowest variability across the expert exclusions. This result indicates that their performance was not driven by any single annotator. Overall, the analysis supports the interpretation that the pairwise framework captures a stable consensus signal rather than overfitting to individual expert preferences. Additional leave-one-expert-out results for sets 2 and 3 are provided in the Supplementary materials.

Outlier and explainability analysis

Figure 7 combines the representative ranking outliers with explainability maps for the same images. Overall, the pairwise model follows the expert-consensus diagonal more closely than the regression model, consistent with the quantitative results. Additional correlation plots for set 1 with pre-trained backbones and for sets 2 and 3 across all backbones, are provided in the Supplementary materials.

The selected examples illustrate two types of large ranking deviation. In one case, the regression model produced a large deviation while the pairwise model remained close to the expert-consensus ranking. In the other case, both models deviated from the consensus; this image has low

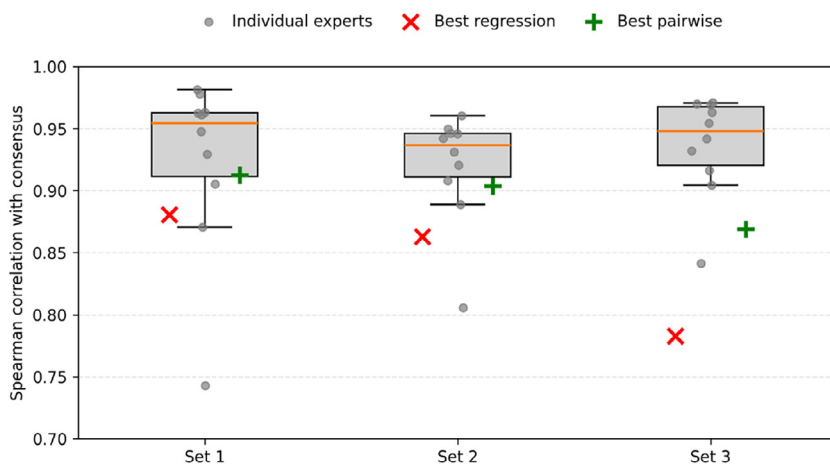


Figure 6. Model rank correlation relative to individual expert variability. For each evaluation set, gray boxplots and points show the distribution of the Spearman correlations between individual expert rankings and the multi-expert consensus ranking. Red crosses indicate the best regression model and green pluses indicate the best pairwise model for each set.

Table 4
Leave-one-expert-out robustness analysis for set 1

Excluded expert	Reg-baseline	Reg-14	Reg-47	Pairwise-baseline	Pairwise-14	Pairwise-47
Expert 1	64.98	88.88	85.05	86.96	91.90	89.05
Expert 2	65.52	88.92	85.27	85.23	91.99	89.54
Expert 3	64.58	88.34	85.81	86.65	91.46	88.97
Expert 4	65.16	88.79	85.49	86.83	91.64	88.88
Expert 5	63.96	87.85	85.09	86.21	91.19	88.88
Expert 6	64.45	88.25	85.23	85.81	91.68	89.05
Expert 7	63.52	89.59	84.78	87.27	91.46	89.23
Expert 8	65.70	88.65	84.56	86.21	91.68	88.70
Expert 9	64.36	88.03	85.14	85.67	91.59	88.88
Expert 10	64.45	88.25	85.23	85.81	91.68	89.05
Mean correlation	64.67	88.56	85.17	86.27	91.63	89.02
Standard variability	0.68	0.51	0.35	0.65	0.23^a	0.23^a

Spearman correlation values are reported for each model after recalculating the reference ranking using a leave-one-expert-out consensus, where one expert was excluded and the rankings of the remaining nine experts were aggregated.

^a Bold values indicate the best-performing result for each model.

visual quality and poorly defined anatomical boundaries, making rotational alignment difficult to assess.

The gradient-weighted class activation mapping (Grad-CAM) and saliency maps provide a qualitative context for these behaviors. In the regression-outlier case, the pairwise visualization appears to be more focused on the central cardiac region, including structures relevant to rotational alignment such as the interventricular septum, ventricular cavities and atrioventricular valve plane. In contrast, the regression activation is more diffuse and extends into less anatomically informative regions. In the shared-outlier case, both models show less clearly localized attention; this finding is consistent with the low image quality and poorly defined endocardial boundaries, which make key landmarks such as the septum and valve plane harder to identify. These visualizations should be interpreted as illustrative rather than definitive evidence of model reasoning, but they suggest that model errors are associated with reduced localization to clinically relevant anatomical cues.

Raw annotation disagreement and cycle consistency

To characterize the consistency of the raw expert annotations before consensus aggregation, each six-image ranking submission was converted into its implied pairwise comparisons. Across 4240 expert ranking submissions, 107,142 pairwise votes were extracted, covering 99,595 unique image pairs. Of these, 7009 pairs were observed more than once, indicating the repeated assessment of the same image pair; 1429 repeated pairs were judged in both directions across the

annotations. The overall minority-vote disagreement rate among repeated pairs was 20.39%. This result indicates that the raw annotation pool contained measurable disagreement before consensus aggregation.

We also assessed the non-transitive structure in the pooled raw expert comparison graph. For each image pair, a majority direction was assigned from the raw pairwise votes and image pairs with exact ties were ignored. We then examined all complete triplets for which all three pairwise majority relations were available. A triplet was considered cyclic if the majority directions implied $i > j$, $j > k$ and $k > i$, or the reverse cycle. The pooled expert majority-vote graph contained 21,204 cyclic triplets among 763,886 complete triplets, corresponding with a raw annotation cycle rate of 2.7758%.

The same cycle-consistency principle was used to assess the raw pairwise predictions of the trained AI models before Elo-style aggregation. Across the three independent external evaluation sets, 12,180 triplets were examined. The baseline and 14-class pre-trained pairwise models produced no cyclic triplets, while the 47-class pre-trained model produced three cyclic triplets, corresponding with a cycle rate of 0.0246%. These results show that the trained pairwise models produced highly stable pairwise preferences on external cases, despite the absence of an explicit transitivity constraint in the architecture. This finding complements the raw annotation analysis, which showed measurable disagreement and a non-transitive structure before the consensus aggregation. However, the human and AI cycle analyses were performed on different graphs and datasets, so these values should be interpreted descriptively rather than as a direct human-vs.-model comparison. Table 5 summaries

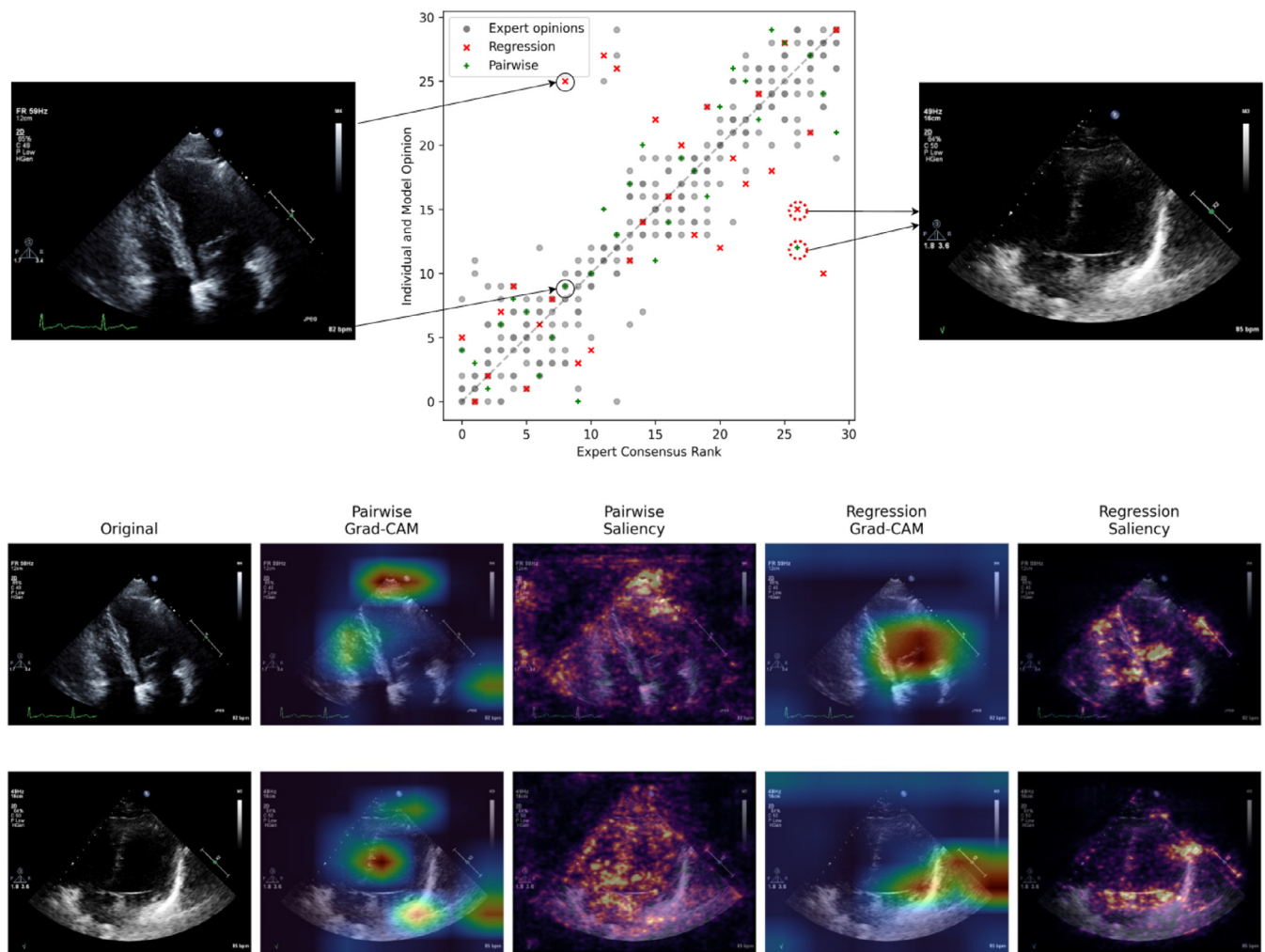


Figure 7. Representative ranking outliers and explainability maps for model-predicted apical four-chamber (A4C) rotational alignment. The *top panel* shows model predictions relative to expert consensus for set-1 using the baseline backbone. *Gray dots* represent the individual expert rankings, *red crosses* represent the regression predictions and *green crosses* represent the pairwise model predictions; points closer to the diagonal indicate a stronger agreement with the expert consensus ranking. The *lower panels* show the gradient-weighted class activation mapping (Grad-CAM) and saliency maps for the selected examples.

Table 5

Cycle consistency of raw expert annotations and trained pairwise model predictions

Source	Complete triplets	Cyclic triplets	Cycle rate
Raw expert majority-vote graph	763,886	21,204	2.7758%
AI pairwise – baseline	12,180	0	0.0000%
AI pairwise – pre-trained-14	12,180	0	0.0000%
AI pairwise – pre-trained-47	12,180	3	0.0246%

AI, artificial intelligence.

Human annotation consistency was assessed on the raw pooled training annotation graph before consensus aggregation. AI consistency was assessed on the raw pairwise predictions from the external evaluation sets before Elo-style aggregation.

the cycle-consistency results for the raw expert majority-vote graph and the trained pairwise models evaluated on the external datasets.

Because the human and AI cycle analyses were performed on different graphs and datasets, these values should not be read as evidence that the AI is directly more consistent than individual experts. Rather, they support the methodological role of the consensus-generation step: raw expert opinions contain disagreement before aggregation, while models trained on consensus-derived pairwise targets can produce internally stable pairwise predictions on external cases.

Conclusion and future work

This study evaluated whether echocardiographic A4C rotational misalignment is better learned as a pointwise regression problem or as a pairwise comparison problem. More broadly, it investigated whether preserving relative expert judgment during model training can improve qualitative echocardiographic rotational alignment assessment. Building on previous work in relative annotation, including multiple-image ranking in cardiac magnetic resonance imaging and our earlier regression-based rotation model, we reformulated the learning stage as a pairwise comparison using targets derived from the expert consensus rotational scale.

Across the three independent expert-ranked evaluation sets, pairwise learning provided a stronger agreement with multi-expert consensus rankings than conventional pointwise regression. This advantage was observed across different backbones and remained stable in leave-one-expert-out analyses, suggesting that the findings were not driven by a single annotator. Grad-CAM and saliency analyses provided complementary qualitative support that the pairwise model focused more strongly on clinically relevant cardiac anatomy, whereas the regression model showed a more diffuse activation.

The human benchmark analysis showed that the model rankings were within the range of individual expert agreement with the consensus, but below the average expert. This finding suggests that the

proposed pairwise formulation captures a clinically meaningful component of expert rotational judgment, while also indicating room for improvement before an expert-level consensus agreement is achieved.

The raw annotation analysis provides further support for the modeling strategy. Raw expert rankings contained measurable pairwise disagreement and a non-transitive structure before the consensus aggregation. Training the pairwise model on consensus-derived score differences was therefore a deliberate design choice: the consensus-generation stage served as a denoising step between the individual expert annotations and pairwise model training. The very low cycle rates observed in the external evaluation sets suggest that this consensus-derived supervision produced internally stable pairwise predictions, despite the absence of an explicit transitivity constraint in the architecture.

These findings support a practical methodological conclusion: for qualitative echocardiographic attributes that are naturally judged by comparison, it may be beneficial to preserve the relative structure during both annotation and model learning. In the specific case of A4C rotational misalignment, pairwise learning offers a more faithful formulation than direct regression from single images to absolute scores. From a clinical perspective, the intended utility is to support a more standardized A4C acquisition, trainee feedback, quality control and the selection of more reliable images for downstream measurement, rather than to replace expert interpretation.

The improved agreement with expert rotational assessment may have practical value in several aspects of echocardiographic acquisition. For sonographer training, automated rotational assessment could provide consistent feedback on clockwise–anti-clockwise alignment and help trainees to recognize subtle deviations from the intended A4C plane. In routine quality control, rotational alignment could complement other image-quality attributes to support a more standardized selection of A4C acquisitions for interpretation or downstream analysis. Greater acquisition consistency may in turn improve the reproducibility of the measurements that depend on standardized chamber geometry, although this factor was not evaluated in the present study.

Several limitations should be acknowledged. The present study evaluated agreement with expert rotational judgment rather than directly quantifying downstream effects on EF, global longitudinal strain, chamber volumes or other clinical measurements. The use of consensus-derived pairwise targets was a deliberate design choice intended to reduce contradictory supervision from noisy individual annotations, although it may discard information about annotator-level uncertainty and disagreement. The current architecture also does not enforce anti-symmetry or transitivity as hard constraints, and the Elo-style aggregation uses the direction of the predicted pairwise differences rather than fully exploiting their continuous magnitude. Computational benchmarking used pre-processed images on a high-end GPU and excluded acquisition and system-integration overhead. Thus, the reported timings indicate model-level feasibility but not end-to-end real-time acquisition guidance.

A further limitation is that rotational alignment represents only one component of multidimensional echocardiographic acquisition. In routine scanning, rotation may coexist with tilt, translation, out-of-plane displacement and foreshortening, all of which can alter the apparent cardiac anatomy and potentially influence model predictions. The present framework should therefore be considered a rotation-specific quality measure rather than a complete acquisition guidance system; future work should evaluate it alongside complementary measures of acquisition quality under prospective scanning conditions.

Future work should compare consensus-derived pairwise supervision with direct training on raw pairwise constraints using probabilistic or uncertainty-weighted ranking losses and should investigate the ranking frameworks that enforce consistency more directly, such as scalar-ranking architectures, bidirectional pair training, antisymmetry regularization, or Bradley–Terry models. Larger multi-center validation is also needed, including prospective assessment within

acquisition workflows and evaluation of whether the proposed rotational index predicts measurement variability or systematic bias in EF, global longitudinal strain, chamber volumes or other clinically derived parameters. Finally, the same pairwise-learning principle should be tested for other qualitative echocardiographic attributes, including endocardial border visibility, anatomical coverage and foreshortening.

Data and code availability

The UNITY-ROT dataset is available through the THRIVE Centre datasets portal: thrive-centre.com/datasets/ROT. Pre-trained and fine-tuned models are released through the [EchoForge model library](#). The source code used for model training, inference and evaluation is available in the [UNITY GitHub repository](#).

Acknowledgements

This research and open-access release were conducted under approvals held by the Imperial College London and the University of West London, London, United Kingdom (IRAS: 279328, REC:20/SC/0386).

Funding

Supported in part by the British Heart Foundation, United Kingdom (Grant no. [RG/F/22/110059](#)). A. Alibakhshi is supported by the Vice Chancellor's Scholarship at the University of West London.

Supplementary materials

Supplementary material associated with this article can be found in the online version at [doi:10.1016/j.ultrasmedbio.2026.09.003](https://doi.org/10.1016/j.ultrasmedbio.2026.09.003).

References

- [1] Dhutia NM, Cole GD, Zolgharni M, Manisty CH, Willson K, Parker KH, et al. Automated speckle tracking algorithm to aid on-axis imaging in echocardiography. *J Med Imaging* 2014;1(3):037001. doi: [10.1117/1.JMI.1.3.037001](#).
- [2] Zaman S, Vimalaevan K, Howard JP, Chappell D, Varela M, Peters NS, et al. Efficient labelling for efficient deep learning: the benefit of a multiple-image-ranking method to generate high volume training data applied to ventricular slice level classification in cardiac MRI. *J Med Artif Intell* 2023;6:4. 5. doi: [10.21037/jmai-22-55](#).
- [3] Fernandes P, Naidoo P, Ufumaka I, Adibzadeh S, Alajrami E, Jevsikov J, et al. Deep learning for assessing rotational misalignment in echocardiographic imaging. *International Conference on AI in Healthcare*. New York: Springer; 2025. p. 269–82.
- [4] Nagata Y, Kado Y, Onoue T, Otani K, Nakazono A, Otsuji Y, et al. Impact of image quality on reliability of the measurements of left ventricular systolic function and global longitudinal strain in 2D echocardiography. *Echo Res Pract* 2018;5:27–39. doi: [10.1530/ERP-17-0047](#).
- [5] Du L, Xue W, Qi Z, Shi Z, Guo G, Yang X, et al. Deep learning framework for interpretable quality control of echocardiography video. *Med Phys* 2025;52:4039–53. doi: [10.1002/mp.17722](#).
- [6] Liu W, Wang Q, Zhang P, Deng Y, Zhao Y, Zhang Y, et al. Automated echocardiogram image quality assessment with yolo and ResNet in the left ventricular myocardium of a4c views. *Appl Intell* 2025;55(6):513.
- [7] Padeloup D, Olaisen SH, Østvik A, Sabo S, Pettersen HN, Holte E, et al. Real-time echocardiography guidance for optimized apical standard views. *Ultrasound Med Biol* 2023;49(1):333–46. doi: [10.1016/j.ultrasmedbio.2022.09.006](#).
- [8] Saebøe S, Pettersen H, Pasdeloup D, Smistad E, Østvik A, Stølen S, et al. Real-time automatic feedback by deep learning to reduce apical foreshortening in echocardiography. *Eur Heart J Cardiovasc Imaging* 2022;23(Supplement_1):jaab289–008.
- [9] Xu L, Li J, Lin W, Zhang Y, Zhang Y, Yan Y. Pairwise comparison and rank learning for image quality assessment. *Displays* 2016;44:21–6. doi: [10.1016/j.displa.2016.06.002](#).
- [10] Liu X, van de Weijer J, Bagdanov AD. Rankiq: learning from rankings for no-reference image quality assessment. In: *Proceedings of the IEEE International Conference on Computer Vision*. Piscataway, NJ: IEEE; 2017. p. 1040–9.
- [11] N. Chahine, S. Ferradans, J. Ponce. Pairwise comparisons are all you need. *arXiv preprint arXiv:2403.09746* (2024).
- [12] Hoeijmakers EJ, Martens B, Hendriks BMF, Muhl C, Miclea RL, Backes WH, et al. How subjective CT image quality assessment becomes surprisingly reliable: pairwise comparisons instead of Likert scale. *Eur Radiol* 2024;34:4494–503. doi: [10.1007/s00330-023-10493-7](#).

- [13] Naidoo P, Fernandes P, Ufumaka I, Dadashi Serej N, Howard J, Francis D, et al. Spatiotemporal contrastive learning for echocardiography view classification. *AI in Healthcare*, Vol. 16039 of Lecture Notes in Computer Science. Cham, Switzerland: Springer Nature;; 2026. p. 247–60. doi: [10.1007/978-3-032-00656-1_18](https://doi.org/10.1007/978-3-032-00656-1_18).
- [14] Lane ES, Azarmehr N, Jevsikov J, Howard JP, Shun-Shin MJ, Cole GD, et al. Multi-beat echocardiographic phase detection using deep neural networks. *Comput Biol Med* 2021;133:104373.
- [15] Naidoo P, Fernandes P, Dadashi Serej N, Stowell CC, Manisty CH, Francis DP, et al. Robust fine-grained echocardiographic view classification with supervised contrastive learning. *Med Image Anal* 2026;110:104006. doi: [10.1016/j.media.2026.104006](https://doi.org/10.1016/j.media.2026.104006).
- [16] Alajrami E, Ng T, Jevsikov J, Naidoo P, Fernandes P, Azarmehr N, et al. Active learning for left ventricle segmentation in echocardiography. *Comp Methods Programs Biomed* 2024;248:108111.