JACC: ADVANCES VOL. 5, NO. 12, 2026
© 2026 THE AUTHORS. PUBLISHED BY ELSEVIER INC ON BEHALF OF THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION. THIS IS AN OPEN ACCESS ARTICLE UNDER THE CC BY LICENSE (http://creativecommons.org/licenses/by/4.0/).

ORIGINAL RESEARCH

Multiview Integration for Robust Echocardiographic Embeddings

Accessible AI for Clinical Practice

Takeshi Tohyama, MD, PhD,[a,b] Ahram Han, MD,[a,c] Dukyong Yoon, MD, PhD,[a,d] Kenneth Paik, MD, MBA,[a] Brian Gow, BS,[a] Nura Izath, BSc, MSc,[e] Leo Anthony Celi, MD, MSc, MPH,[a,f,g] Jacques Kpodonu, MD[h]

ABSTRACT

BACKGROUND Although recent artificial intelligence (AI) foundation models have shown superior performance in medical imaging, their high computational demands and reliance on large data sets limit accessibility. These issues are especially pronounced in echocardiography, where integrating complex multiview measurements for individual cases hinders AI development and clinical translation. Vector embedding approaches offer promising solutions by leveraging compact representations in AI applications.

OBJECTIVES We aimed to develop a transformer-based multiview encoder that integrates multiple echocardiographic videos into unified vector embeddings and uses adversarial learning to reduce reliance on demographic shortcuts.

METHODS Using the MIMIC-IV-ECHO (Medical Information Mart for Intensive Care IV-Echocardiography) data set (7,169 studies), we trained a multiview encoder to produce study-level embeddings and evaluated them on 21 binary classification tasks covering echocardiographic measurements and clinical diagnoses. We varied adversarial weights to assess their impact on demographic shortcut reliance and overall performance.

RESULTS The multiview encoder raised the mean area under the curve (AUC) across clinical tasks from 0.742 to 0.831 compared with the foundation model embedding (ΔAUC +0.090; 95% CI: 0.078-0.100), with significant gains in 20 of 21 tasks. The embeddings remained robust with a limited number of views (mean ΔAUC +0.071; 95% CI: 0.043-0.113). Adversarial learning reduced sex and race predictability only at strong weights and at substantial cost to clinical performance.

CONCLUSIONS The multiview encoder framework integrates multiple echocardiographic views into unified study-level embeddings, enabling substantial diagnostic improvements across diverse cardiac tasks. By producing compact, informative embeddings without requiring extensive computational resources, this approach lowers the cost of building echocardiographic AI on an existing foundation model. (JACC Adv. 2026;5:103488) © 2026 The Authors. Published by Elsevier Inc on behalf of the American College of Cardiology Foundation. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

From the [a]Institute for Medical Engineering and Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; [b]International University of Health and Welfare, Fukuoka, Japan; [c]Department of Surgery, Seoul National University Hospital, Seoul, Republic of Korea; [d]Department of Biomedical Systems Informatics, Yonsei University College of Medicine, Seoul, Republic of Korea; [e]Faculty of Computing and Informatics, Mbarara University of Science and Technology Data Science Research Hub (MUDSReH), Mbarara, Uganda; [f]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA; [g]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA; and the [h]Division of Cardiac Surgery, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA.

ISSN 2772-963X https://doi.org/10.1016/j.jacadv.2026.103488

ABBREVIATIONS AND ACRONYMS

AI = artificial intelligence

AUC = area under the curve

DICOM = Digital Imaging and Communications in Medicine

GPU = graphics processing unit

ICD = International Classification of Diseases

LVEF = left ventricular ejection fraction

MIMIC = Medical Information Mart for Intensive Care

MVE = multiview encoder

RV = right ventricular

VE = vector embedding

Echocardiography serves as the cornerstone of cardiovascular diagnostics by providing real-time visualization of cardiac structure and function through multiple standardized imaging views.[1] Although recent foundation models, such as EchoPrime and PanEcho, have demonstrated superior capabilities across diverse cardiac imaging tasks,[2-5] their massive computational requirements and reliance on large-scale proprietary datasets create substantial barriers, limiting artificial intelligence (AI) development and clinical translation to institutions with extensive resources.[6,7] The inherent complexity of echocardiographic assessment, which requires comprehensive multiview evaluation and integration of diverse videos, including color Doppler, exacerbates these barriers.[8,9]

Vector embedding (VE) offers a promising alternative approach that can substantially reduce these barriers by leveraging the compact representations extracted from pretrained models for diverse downstream applications. Recent studies have demonstrated that embedding-based approaches achieve performance comparable to that of larger foundation models, particularly in data-limited settings.[10,11] Furthermore, downstream AI models can be efficiently trained using embeddings on affordable hardware in hours rather than weeks, enabling resource-limited institutions to develop dedicated models without requiring massive computational infrastructure or large-scale data sets.

In addition to these accessibility concerns, demographic fairness remains a critical challenge in medical AI. Models may unintentionally learn spurious demographic shortcuts rather than clinically relevant pathophysiological features, thereby creating potential biases that can lead to disparate performance across demographic groups.[12] Although adversarial learning approaches have been proposed to mitigate these biases,[13,14] their application and effectiveness, specifically in echocardiography, remain unclear.

Herein, we propose a multiview encoder (MVE) framework that addresses the computational accessibility and demographic shortcut challenges in current echocardiographic AI. The framework integrates information across multiple imaging views through compact VE representations, thereby reducing the computational requirements while improving diagnostic performance. We also examined, through adversarial learning, how much demographic information the embeddings retain and whether reducing it affects performance across demographic groups.

METHODS

STUDY DESIGN AND DATA SOURCE. This retrospective study used the Medical Information Mart for Intensive Care IV (MIMIC-IV) and MIMIC-IV-ECHO datasets.[15-17] These publicly available data sets contain anonymized echocardiographic examinations from Beth Israel Deaconess Medical Center. From the MIMIC-IV-ECHO data set, all patients with clinical diagnostic information between 2017 and 2019 were included, with the analysis limited to echocardiographic examinations containing both Digital Imaging and Communications in Medicine (DICOM) video files and corresponding measurement records. The Institutional Review Board of the Beth Israel Deaconess Medical Center approved the creation of these databases and granted a waiver of informed consent; our secondary use required no additional approval because the records were already deidentified in accordance with the Health Insurance Portability and Accountability Act Safe Harbor provisions. All analyses were conducted in accordance with the Declaration of Helsinki, and this study is reported in accordance with the TRIPOD+AI statement and PRIME 1.0 guidelines for cardiovascular imaging-related machine learning evaluation.[18,19] The complete PRIME 1.0 checklist is presented in Supplemental Table 1. No prospective protocol registration was conducted because this retrospective analysis used publicly available data sets. No patient or public involvement was undertaken in this retrospective study.

OUTCOME DEFINITION FOR CLINICAL TASKS. We defined 21 classification tasks, encompassing echocardiographic measurements, functional assessments, and clinical diagnoses (**Table 1**). The tasks were defined for this study. Primary outcomes

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the Author Center.

Manuscript received January 6, 2026; revised manuscript received October 14, 2026, accepted October 14, 2026.

TABLE 1 Characteristics of the Echocardiographic Studies, Overall and by Data Split

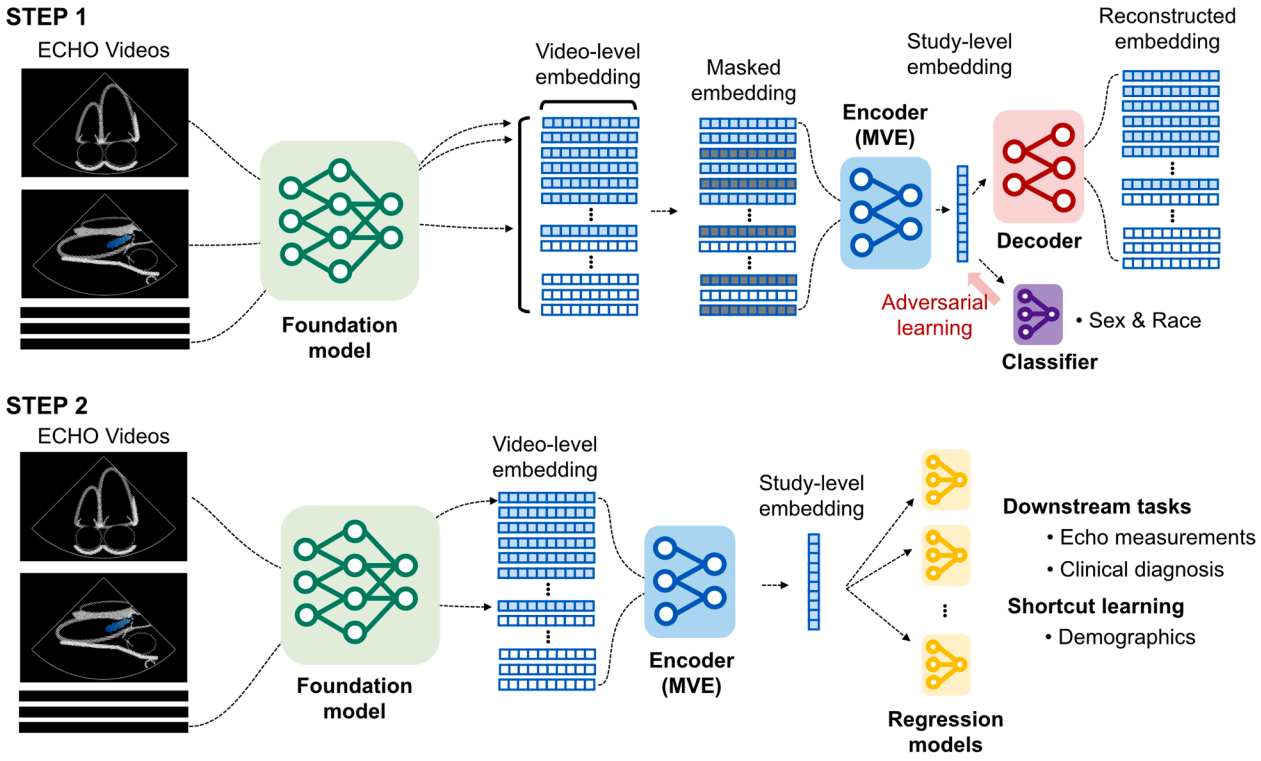
	Missing, n (%)	Overall	MVE Training	MVE Validation	Evaluation
Studies, n		7,169	3,033 (2,990-3,088)	531 (503-557)	3,605 (3,554-3,635)
Patients, n		4,555	1,937	341	2,277
Video count, median		41	41 (40-41)	41 (41-42)	41
Age, y, mean	0 (0.0)	70.7	70.6 (70.3-71.0)	70.8 (70.5-71.2)	70.7 (70.3-71.0)
Sex, %: Female	417 (5.8)	51.5	51.2 (50.6-52.5)	51.0 (45.3-55.3)	51.8 (50.4-53.1)
Sex, %: Male		48.5	48.8 (47.5-49.4)	49.0 (44.7-54.7)	48.2 (46.9-49.6)
Race, %: Black	498 (6.9)	18.5	18.3 (17.1-19.1)	18.1 (14.4-19.7)	18.6 (18.1-19.6)
Race, %: Others		9.9	9.8 (9.4-10.1)	9.2 (6.8-11.6)	10.1 (9.9-10.5)
Race, %: White		71.6	71.8 (70.9-73.5)	72.7 (69.9-74.9)	71.3 (70.3-71.9)
Aortic root/BSA >2.1 cm/m^2, %	3,153 (44.0)	6.6	6.3 (5.9-6.7)	4.7 (3.7-5.1)	7.1 (6.7-7.6)
LAVI >34 mL/m^2, %	2,711 (37.8)	41.3	40.8 (39.8-42.2)	41.9 (38.9-45.0)	41.8 (41.0-42.1)
LVMI >115 g/m^2 (M) or 95 g/m^2 (F), %	5,129 (71.5)	32.0	31.7 (30.4-33.0)	30.0 (25.2-33.0)	32.6 (31.5-33.7)
LVDd >5.8 cm (M) or 5.2 cm (F), %	2,542 (35.5)	10.5	10.6 (9.9-11.9)	10.2 (7.3-13.4)	10.6 (8.9-11.1)
LVEF <40%, %	1,226 (17.1)	10.6	10.6 (9.8-11.6)	8.7 (5.6-12.8)	10.8 (9.3-11.4)
E/e′ >15, %	3,576 (49.9)	19.5	19.6 (18.5-20.5)	18.8 (14.9-21.4)	19.5 (18.7-20.3)
TRPG >31 mm Hg, %	1,743 (24.3)	34.5	33.3 (32.1-35.5)	36.1 (31.2-40.3)	35.2 (32.7-36.6)
Impaired RV function, %	2,328 (32.5)	16.0	15.4 (14.2-17.2)	16.6 (11.9-21.0)	16.3 (14.5-17.4)
≥Mild AS, %	143 (2.0)	7.5	7.5 (7.4-7.7)	7.6 (6.5-8.6)	7.5 (7.3-7.6)
≥Mild AR, %	139 (1.9)	15.7	15.0 (14.5-15.6)	16.1 (13.2-18.0)	16.3 (16.0-16.5)
≥Mild MR, %	131 (1.8)	34.4	34.3 (33.6-35.7)	36.3 (33.3-40.2)	34.2 (33.2-34.7)
≥Mild TR, %	128 (1.8)	36.3	36.1 (35.7-36.7)	36.7 (32.8-38.5)	36.4 (36.0-37.2)
≥Moderate AS, %	143 (2.0)	4.3	4.4 (4.1-4.8)	4.1 (3.2-5.4)	4.3 (4.1-4.6)
≥Moderate AR, %	139 (1.9)	1.6	1.6 (1.3-2.2)	1.5 (0.8-2.4)	1.6 (1.3-1.9)
≥Moderate MR, %	131 (1.8)	9.0	8.9 (8.4-9.6)	8.4 (6.3-11.6)	9.1 (8.7-9.5)
≥Moderate TR, %	128 (1.8)	10.4	9.9 (9.5-10.4)	10.8 (8.7-13.4)	10.8 (10.6-11.0)
Dilated IVC, %	4,494 (62.7)	19.6	19.4 (18.4-20.9)	17.0 (13.7-20.6)	20.1 (18.3-21.2)
Myocardial infarction, %	3,647 (50.9)	5.8	6.0 (5.3-6.6)	3.9 (2.5-6.6)	5.8 (4.9-6.4)
Pulmonary embolism, %	2,991 (41.7)	1.5	1.1 (0.6-1.8)	1.7 (1.0-2.3)	1.8 (1.3-2.1)
Dilated cardiomyopathy, %	361 (5.0)	6.7	6.6 (5.6-7.7)	6.6 (5.3-7.5)	6.8 (5.7-7.8)
Hypertrophic cardiomyopathy, %	361 (5.0)	1.8	1.9 (1.6-2.4)	1.1 (0.2-2.0)	1.8 (1.4-2.1)

Percentages are of studies with the variable recorded, and the Missing column refers to the overall cohort. Values in the 3 split columns are the mean across the 5 repeats, with the range in parentheses.

AR = aortic regurgitation; AS = aortic stenosis; BSA = body surface area; E/e′ = ratio of early mitral inflow velocity to early diastolic mitral annular tissue velocity; IVC = inferior vena cava; LAVI = left atrial volume index; LVDd = left ventricular diameter (diastolic); LVEF = left ventricular ejection fraction; LVMI = left ventricular mass index; MR = mitral regurgitation; MVE = multiview encoder; RV = right ventricular; TR = tricuspid regurgitation; TRPG = tricuspid regurgitation pressure gradient.

included echo measurements (eg, reduced ejection fraction, right ventricular (RV) dysfunction, and valvular disease) and clinical diagnoses such as cardiomyopathy and myocardial infarction. No additional outcome assessments were conducted. All labels were taken from existing clinical records; no free-text natural language processing and no manual rereading of images were performed. The 17 measurement and functional tasks came from the structured measurements distributed with MIMIC-IV-ECHO, values committed by the interpreting clinician as part of routine care rather than a research reread; each value is therefore a routine clinical read, although the databases do not record whether a second reader reviewed it. All echo measurements were dichotomized using clinically validated thresholds established in the current guidelines.[20,21] Valvular tasks are named with an inequality (eg, ≥mild mitral regurgitation, ≥moderate mitral regurgitation) to show that they denote mild-or-greater or moderate-or-greater severity, not a single grade. Not every study had a recorded valvular grade. A study imaged with color Doppler but without a recorded grade was labeled negative. Supplemental Table 2 reports the proportion of negatives assigned by this rule for each valvular task. The 4 diagnosis tasks came from International Classification of Diseases (ICD)-9/10 codes in the MIMIC-IV hospital records; a positive label for myocardial infarction and pulmonary embolism additionally required a concordant echocardiographic finding (reduced ejection fraction and tricuspid gradient >31 mm Hg,

FIGURE 1 Two-Step Framework for Multiview Echocardiographic Representation Learning



Step 1 (pretraining phase): The foundation model generates video-level embeddings from the echocardiographic videos. The MVE integrates these embeddings into study-level representations. Adversarial learning simultaneously suppresses demographic shortcuts (sex and race) in the embedding. Step 2 (evaluation phase): The vector embeddings produced by the MVE are assessed using regression models, evaluating performance across 21 downstream clinical tasks and examining robustness against demographic shortcut learning. MVE = multiview encoder.

respectively), because the code alone does not distinguish an old event from an acute one; a negative label required only that the code be absent, and a study whose code and finding disagreed was left unlabeled. Full task definitions, the derived measurements, and the ICD code lists are given in the Supplemental Appendix.

Missing labels were handled by task-level complete-case analysis. Encoder pretraining did not use labels and was therefore unaffected. A study without a recorded label was used neither for fitting nor for evaluating that task's downstream classifier, and no values were imputed.

VIDEO-LEVEL VE EXTRACTION FROM FOUNDATION MODEL. For each echocardiographic study, we first extracted 512-dimensional video-level VEs from every video using the EchoPrime foundation model, which was developed on more than 12 million video-report pairs via contrastive learning at Cedars-Sinai Medical Center; MIMIC-IV-ECHO was not included in its training data.[3] MIMIC-IV-ECHO contained a median of 41 videos per study (maximum: 118), representing multiple standardized echocardiographic videos, including parasternal, apical, and subcostal views. The MVE then used the full set of per-video vectors as input to represent the entire study. All VE extraction was automated, and investigators did not manually derive features or visually assess the images.

MVE MODEL DEVELOPMENT. The model uses a transformer-based architecture with 2 transformer layers and 8 attention heads optimized for computational efficiency using compact transformer approaches.[22-24] This architecture simultaneously processes a variable number of cardiac videos (Figure 1). Each 512-dimensional video embedding is linearly projected and then passes through the shared encoder independently, and the embeddings of a study are integrated by an attention aggregator with a global query. Pretraining is self-supervised:

half the view embeddings are replaced by a learnable mask token and a linear decoder reconstructs them from the rest, minimizing mean squared error at the masked positions. The MVE aggregates variable numbers of echo video embeddings per study using attention-masking mechanisms, supporting up to 128 videos as inputs to the model.

To ensure that the learned representations captured clinically relevant features rather than spurious demographic shortcuts, an adversarial learning component with gradient reversal layers was integrated during MVE pretraining.[13,25] The adversarial framework included separate discriminator networks for sex prediction (3 classes: female/male/unknown) and race prediction (4 classes: White/Black/Others/Unknown), trained simultaneously with clinical tasks. The gradient reversal layer applies negated gradients (scaled by the adversarial weight ω) during backpropagation, forcing the encoder to learn representations with lower predictability of demographic attributes while maintaining performance in clinical tasks. We systematically varied the adversarial weight (ω) from 0 to 1.0 (0, 0.1, 0.2, 0.5, 1.0) to investigate trade-offs between model performance and demographic predictability. An ω of 0 indicates no adversarial learning, whereas an ω of 1 represents strong adversarial learning.

Our study consisted of 2 steps: 1) pretraining the MVE using random video masking for multiview representation learning; and 2) evaluating the MVE VE on downstream tasks. To prevent patient overlap, we partitioned the data by subject into development (50%) and evaluation (50%) subsets (Supplemental Figure 1). The development subset was divided further by subject into MVE training (85%) and validation (15%). Both partitions were redrawn for each of the 5 repeats from that repeat's seed. The MVE with adversarial learning components was trained using the development subset. The trained MVE was then applied to video-level embeddings in the evaluation subset to generate study-level embeddings. The MVE was trained for 50 epochs using the AdamW optimizer with learning rate 1×10^{-4} and early stopping based on validation loss. Architectural parameters were selected based on best practices from previous studies and preliminary experiments.

DOWNSTREAM TASK PERFORMANCE EVALUATION. We used subject-based 4-fold cross-validation on the evaluation subset for downstream task evaluation, repeating the entire procedure 5 times with different random seeds to ensure robust performance estimates (Supplemental Figure 1). Each repeat was leakage-free: its MVE was trained on that repeat's development subset and evaluated only on its evaluation subset, so no patient contributed to both encoder training and its evaluation. For each study, we averaged the out-of-fold predicted probabilities across the repeats in which it was held out, yielding 1 score per study. For the downstream classification tasks, we used a simple logistic regression model with an L2 penalty whose strength was selected for each representation by subject-level inner cross-validation within the training split of every fold, with the held-out fold taking no part in the selection. The primary evaluation was based on the area under the receiver-operating characteristic curve and area under the curve (AUC). No specific class-imbalance method was used. To provide a comprehensive evaluation, we supplemented the AUC with accuracy and the F1 score, where the decision threshold was determined by optimizing F1 on the training data.

We compared the study-level VEs on downstream task performance. The Foundation VE served as the baseline and consisted of embeddings directly extracted from EchoPrime using the original data-processing pipeline. This approach aggregates variable-length input sequences into fixed dimensions of 15×512 through weighted averaging based on view classification. MVE VEs include embeddings extracted from MVEs with varying adversarial weights, with each embedding having 512 dimensions.

DEMOGRAPHIC PREDICTABILITY AND IMPACT ON CLINICAL TASKS. We systematically evaluated potential demographic shortcuts by training separate classifiers to predict sex and race from the Foundation VE and MVE VE representations. We assessed demographic predictability to evaluate shortcut learning.[26,27] This approach provides one of the metrics for evaluating whether adversarial learning has reduced reliance on demographic shortcuts in VE representations. It measures the demographic information recoverable from an embedding rather than fairness itself.

In addition, we evaluated the effects of adversarial learning weighting for 4 cardiac conditions with sufficient prevalence (≥Mild aortic stenosis [AS], left ventricular ejection fraction [LVEF] <40%, impaired RV function, and E/e′ ratio >15) across demographic subgroups (sex: female/male; race: White/Black/Others). Furthermore, we assessed how adversarial learning affected actual clinical task prediction performance. To evaluate potential disparities in model performance, we compared AUC values on 21 clinical tasks across demographic subgroups (sex: female/male; race: White/Black/Others) and video-count

subgroups (<20, 20-39, 40-59, and ≥60 videos per study). The video-count subgroups reflect the number of videos acquired in each examination rather than views removed experimentally.

COMPUTATIONAL COST ASSESSMENT. To quantify accessibility, we measured computational cost on 1 workstation graphics processing unit (GPU). We timed training of the MVE and separately timed running the EchoPrime pipeline over a study's DICOM files, on 100 randomly drawn studies. Hardware, software versions, and per-stage timings are given in the Supplemental Appendix.

STATISTICAL ANALYSIS. This study used the entire MIMIC-IV-ECHO data set (n = 7,169) for comprehensive analyses rather than individual sample size calculations. This approach was justified for the following reasons: 1) the comprehensive evaluation across 21 diverse clinical tasks with varying prevalence rates (1.6%-41.6%) (Supplemental Table 3) would require different sample sizes for each task, making a single calculation impractical; and 2) using all available data maximizes statistical power and ensures representativeness of the source population.

All uncertainty was quantified with a patient-level (subject cluster) bootstrap, with the patient as the unit of resampling and the study as the unit of scoring. For each embedding, we first stabilized the point estimate by averaging, per study, the out-of-fold predicted probabilities across the repeats in which that study was held out. We then drew 2,000 bootstrap replicates, resampling patients with replacement stratified by outcome, including all studies of each resampled patient, and computing study-level AUC on the resampled studies. The same resample was applied to every embedding within a replicate, so between-embedding differences (ΔAUC = AUC[MVE] − AUC[Foundation]) form a paired comparison. CIs are the 2.5th and 97.5th percentiles of the bootstrap distribution, and two-sided *P* values are twice the proportion of replicates falling on the opposite side of the null. The mean improvement across the 21 tasks was the mean per-task ΔAUC within each replicate, using one common unstratified resample across tasks, so the CI accounts for patients shared across tasks. Subgroup interaction was assessed as a difference-in-differences: the MVE advantage in one subgroup minus that in a reference subgroup, with contrasts for sex (male vs female), race (Black vs White and Others vs White), and video count (<20, 20-39, and 40-59 vs ≥60), summarized per task and averaged. We did not apply multiple-testing corrections. All analyses used Python 3.11 with NumPy 2.4, pandas 2.3, scikit-learn 1.8, and PyTorch 2.6 (CUDA 12.6).

RESULTS

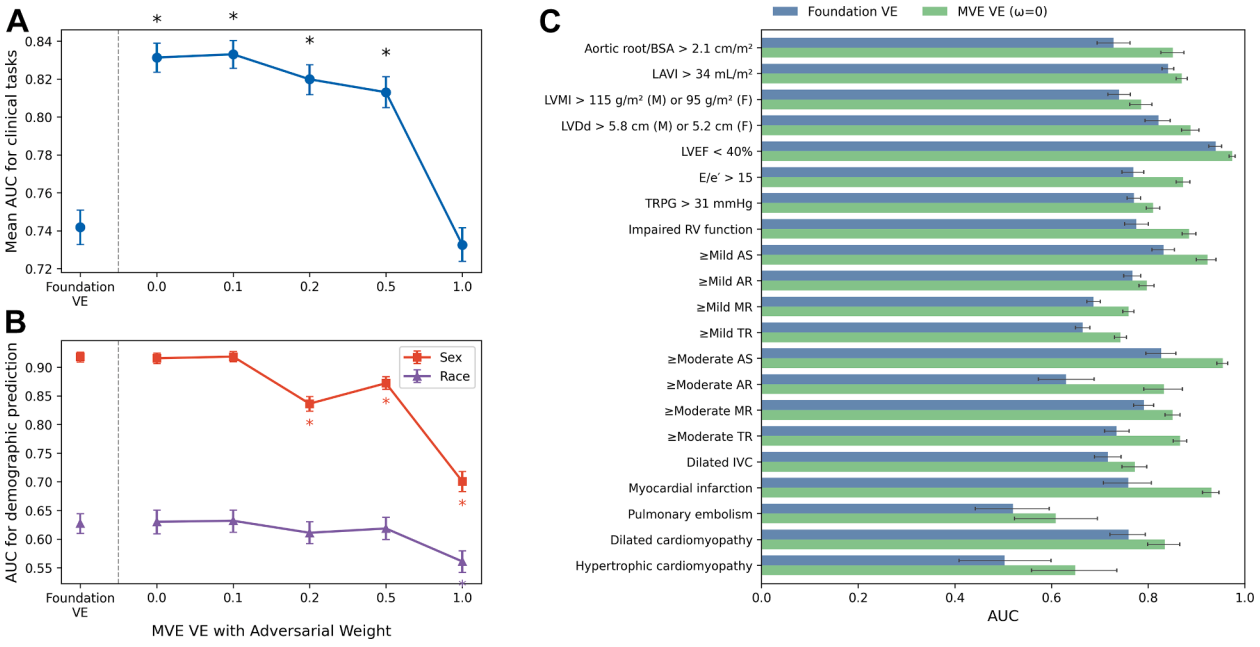
DATA SET CHARACTERISTICS. Our analysis included 7,169 echocardiographic studies, with a median of 41 videos per study (IQR: 34-48) (**Table 1**). The studies came from 4,555 patients (mean age 70.7 years). Among studies with recorded sex (n = 6,752), 3,478 (51.5%) were from female and 3,274 (48.5%) from male patients. Among studies with recorded race (n = 6,671), the racial distribution comprised White (71.6%, n = 4,777), Black (18.5%, n = 1,232), and Others (9.9%, n = 662).

Because the partition was redrawn for each repeat, 6,946 of the 7,169 studies were held out at least once; this combined evaluation set underlies the per-task counts in Supplemental Table 3, whereas **Table 1** describes the whole cohort. The subject-level partition produced comparable splits (**Table 1**): MVE training (1,937 patients, mean 3,033 studies), validation (341 patients, 531 studies), and evaluation (2,277 patients, 3,605 studies). Averaged over the 5 repeats, sex, race, video count, and the prevalences of the measurement-derived tasks were similarly distributed across the splits.

Echocardiographic measurements covered diverse cardiac conditions with varying prevalence, including reduced ejection fraction (LVEF <40%, 10.6%), impaired RV function (16.0%), mild or greater mitral regurgitation (34.4%), mild or greater tricuspid regurgitation (36.3%), elevated tricuspid regurgitation pressure gradient (>31 mm Hg, 34.5%), and elevated E/e′ ratio (>15, 19.5%). The clinical diagnoses included hypertrophic cardiomyopathy (1.8%), dilated cardiomyopathy (6.7%), myocardial infarction (5.8%), and pulmonary embolism (1.5%).

PERFORMANCE COMPARISON OF VECTOR EMBEDDINGS ON DOWNSTREAM TASKS. To establish benchmark MVE VE performance on downstream tasks, we first evaluated MVE VE without adversarial learning (ω = 0) across 21 binary clinical tasks spanning valvular disease, functional assessment, hemodynamic parameters, and clinical diagnoses. Averaged across the 21 tasks, the AUC was 0.831 with the MVE VE and 0.742 with the Foundation VE, a mean ΔAUC of +0.090 (95% CI: 0.078-0.100) (**Figure 2A**). Without adversarial learning, sex remained highly predictable from the MVE VE (0.916 vs 0.918 for the Foundation VE), and race predictability was far lower and likewise similar (0.630 vs 0.627).

FIGURE 2 Clinical Performance and Demographic Predictability



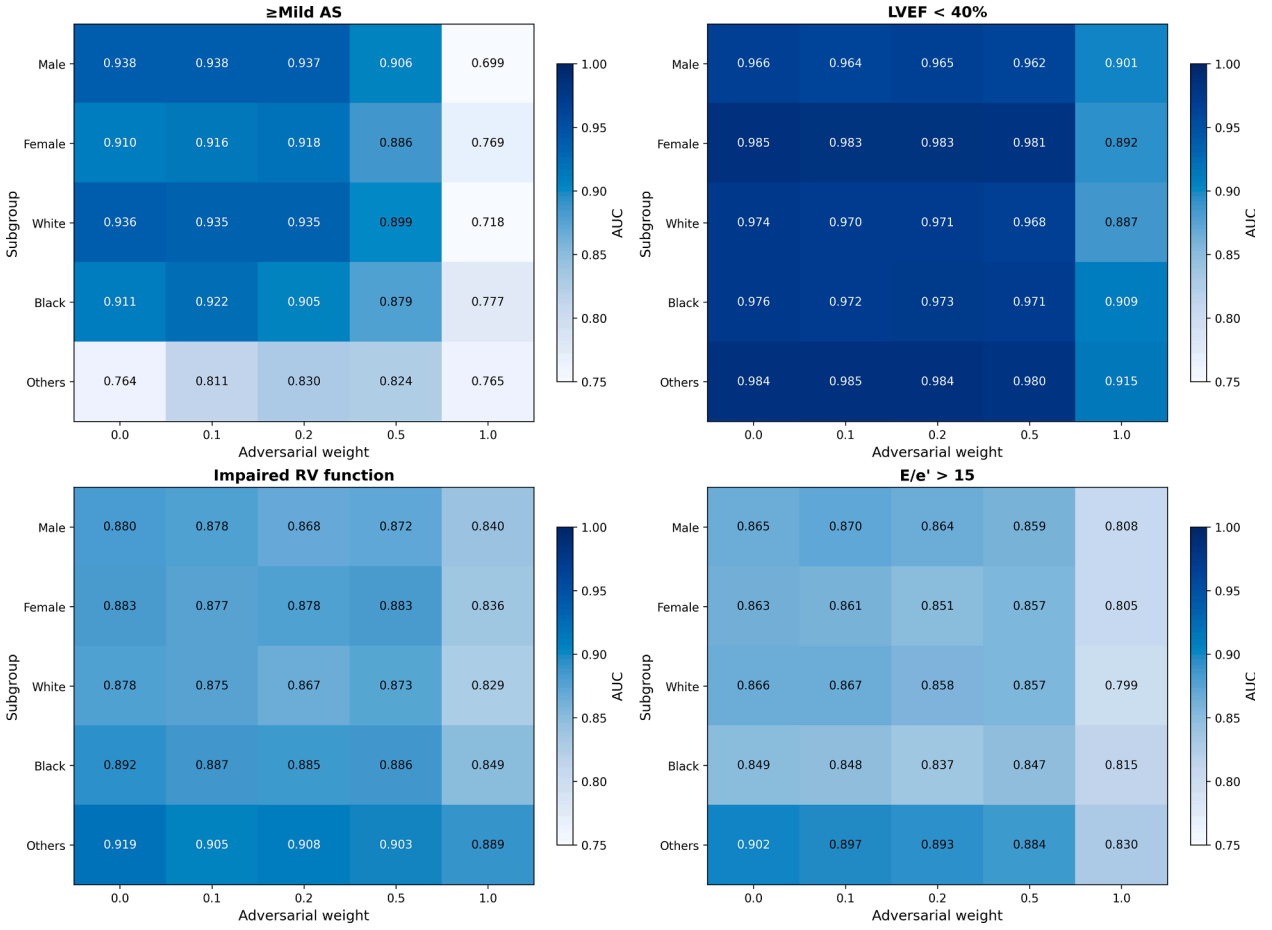
(A) Mean AUC across the 21 clinical tasks for the Foundation VE and for MVE VE trained with adversarial weight ω = 0 to 1.0. Points are the mean across tasks; whiskers are patient-level bootstrap 95% CIs; asterisks mark models whose mean ΔAUC vs the Foundation VE has a 95% CI excluding zero. (B) AUC for predicting sex (red squares) and race (purple triangles) from each embedding (lower values indicate less recoverable demographic information), with 95% CIs; asterisks mark models whose reduction vs the Foundation VE has a 95% CI excluding zero. (C) Per-task AUC for the Foundation VE and the primary MVE VE (ω = 0), with 95% CIs. Point estimates improved for all 21 tasks; 20 intervals excluded zero. AR = aortic regurgitation; AS = aortic stenosis; AUC = area under the curve; BSA = body surface area; E/e = ratio of early transmitral flow velocity to early diastolic mitral annular velocity; IVC = inferior vena cava; LAVI = left atrial volume index; LVDd = left ventricular end-diastolic dimension; LVEF = left ventricular ejection fraction; LVMI = left ventricular mass index; MR = mitral regurgitation; MVE = multi-view encoder; RV = right ventricle; TR = tricuspid regurgitation; TRPG = tricuspid regurgitation pressure gradient; VE = vector embedding.

To mitigate the demographic shortcuts, we systematically varied the adversarial weight (ω) from 0.1 to 1.0 to evaluate its effect on both downstream task performance and demographic predictability (**Figures 2A and 2B**). The MVE VE with mild adversarial learning (ω = 0.1) achieved a mean AUC of 0.833 (mean ΔAUC +0.091 over the Foundation VE; 95% CI: 0.081-0.101), similar to that without adversarial learning, and its demographic predictability was also little changed from ω = 0. Further increases in adversarial weighting reduced sex predictability (0.916 at ω = 0 to 0.701 at ω = 1.0); however, this was accompanied by a substantial performance degradation in clinical tasks (**Figure 2A**). Race predictability also fell, from 0.630 at ω = 0 to 0.561 at ω = 1.0, but by less than sex predictability (0.069 vs 0.215) and from a much lower starting level.

Across individual clinical tasks, point estimates improved in all 21 tasks, and 20 had a 95% CI excluding zero, with a median ΔAUC of +0.078 (IQR: 0.056-0.123) using the MVE VE (ω = 0) compared with the Foundation VE (**Figure 2C**); the only interval spanning zero was pulmonary embolism, the task with the fewest positive cases. F1 improved in all 21 tasks (median +0.091; IQR: 0.056-0.123), whereas accuracy, a threshold-dependent measure dominated by the majority class, changed little (median +0.003; IQR: −0.001 to 0.026) (Supplemental Figure 2).

Tasks differed markedly in difficulty and in precision (Supplemental Table 3). Discrimination with the MVE VE ranged from 0.974 (95% CI: 0.968-0.980) for LVEF <40% and 0.931 (0.913-0.946) for myocardial infarction down to 0.609 (0.523-0.695) for pulmonary embolism and 0.649 (0.559-0.735) for hypertrophic cardiomyopathy, and CI width tracked the number of labeled cases. The least well discriminated tasks were 2 of the ICD-coded diagnoses with the fewest positive cases, pulmonary embolism (prevalence 1.6%) and hypertrophic cardiomyopathy (1.8%); for hypertrophic cardiomyopathy, the Foundation VE was at chance (0.503; 95% CI: 0.408-0.599) whereas the MVE VE reached 0.649 (0.559-0.735).

FIGURE 3 Adversarial Learning Effects Across Demographic Subgroups

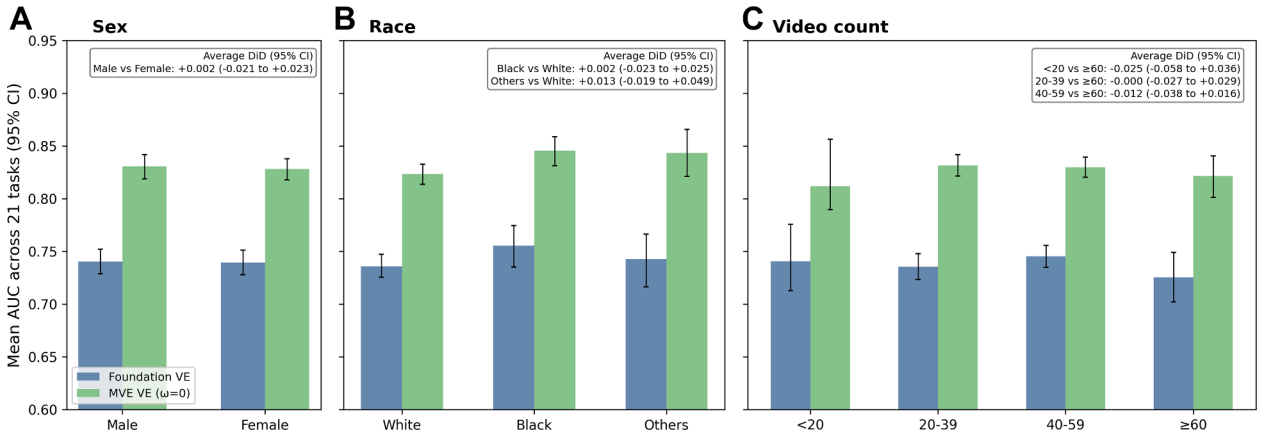


Heatmap showing area under the curve (AUC) performance by demographic subgroups and adversarial weighting (ω = 0-1.0) for 4 clinical tasks: mild or greater aortic stenosis (AS), reduced ejection fraction (LVEF <40%), impaired right ventricular (RV) function, and E/e′ >15. Color intensity represents AUC values, with darker colors indicating higher AUC performance. Each row represents a different demographic subgroup (Male, Female, White, Black, and Others), and each column represents adversarial weights from 0 to 1.0. AS = aortic stenosis; AUC = area under the curve; E/e = ratio of early mitral inflow velocity to early diastolic mitral annular velocity; LVEF = left ventricular ejection fraction; RV = right ventricle.

PERFORMANCE DISPARITIES ACROSS DEMOGRAPHIC GROUPS. To further investigate task-level performance disparities across subgroups, we analyzed 4 specific cardiac outcomes (Figure 3). Among these, reduced ejection fraction (LVEF <40%), impaired RV function, and elevated E/e′ ratio showed only modest subgroup differences at ω = 0. In contrast, ≥Mild aortic stenosis showed lower performance in the Others race category (AUC 0.764 vs 0.936 in White patients at ω = 0). Mild adversarial weights narrowed this gap only partially (0.830 vs 0.935 at ω = 0.2), and stronger weights did not close it further without degrading performance; at ω = 1.0, discrimination fell substantially across all tasks and subgroups. In this setting, adversarial learning reduced demographic predictability only at a substantial cost to clinical performance.

Performance analysis across demographic subgroups revealed that the MVE VE consistently outperformed the Foundation VE across all evaluated subgroups (Figures 4A and 4B). Averaged across the 21 tasks, the MVE advantage did not differ between subgroups: the difference-in-differences was +0.002 (95% CI: −0.021 to 0.023) for male vs female, +0.002 (−0.023 to 0.025) for Black vs White, and +0.013 (−0.019 to 0.049) for Others vs White, every interval

FIGURE 4 Performance Across Demographic and Video-Count Subgroups



Performance comparison across 21 clinical tasks showing area under the curve (AUC) with 95% CIs for foundation model embeddings (Foundation VE) vs multiview encoder embeddings (MVE VE). MVE was trained without adversarial learning. Data are stratified by demographic subgroups: sex (A: female vs male), race (B: White vs Black vs Others), and video count per study (C: <20, 20-39, 40-59, and ≥60). AUC = area under the curve; MVE = multiview encoder; VE = vector embedding.

spanning zero. Performance improvements were consistent across all demographic groups, without widening the existing subgroup differences.

PERFORMANCE WITH LIMITED NUMBERS OF ECHOCARDIOGRAPHIC VIEWS. The MVE improvement was present at every level of examination completeness: mean ΔAUC +0.071 (95% CI: 0.043-0.113) for studies with fewer than 20 videos, +0.096 (0.082-0.110) for 20-39, +0.084 (0.071-0.097) for 40-59, and +0.096 (0.072-0.119) for 60 or more (**Figure 4C**).

COMPUTATIONAL COST. Our framework requires running the EchoPrime pipeline over each study's DICOM files to obtain per-video embeddings, which takes a median of 40.0 seconds per study, most of which is spent reading and decoding those files (Supplemental Appendix). Training the MVE, which aggregates these embeddings into a single 512-dimensional study vector, took 204 seconds on a single workstation GPU (7.9 million parameters; 0.52 GB peak GPU memory). Once that vector is stored, the 21 predictions need only a logistic regression applied to 2 KB of data on a central processing unit, so a single stored vector serves every further task at negligible cost. Supplemental Table 4 gives the details and a comparison with EchoPrime.

DISCUSSION

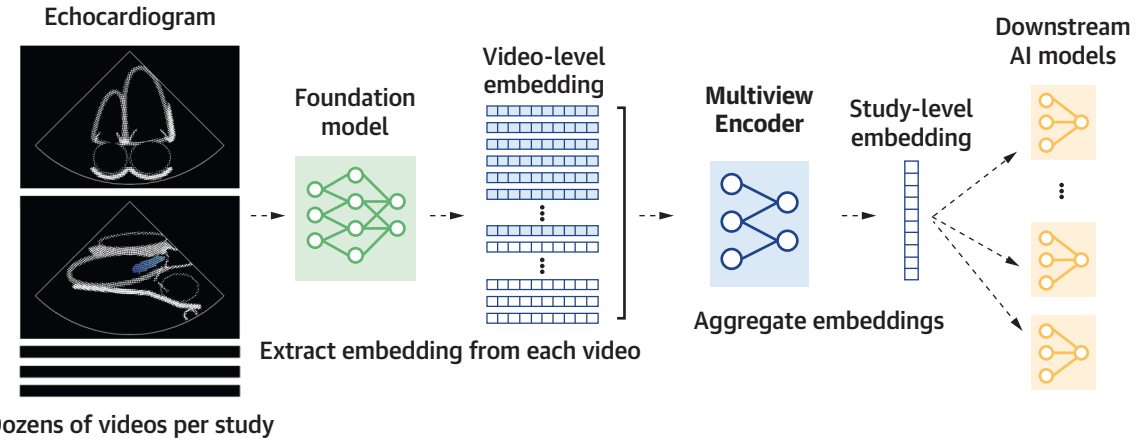
KEY FINDINGS AND CLINICAL SIGNIFICANCE. Our study demonstrated that the MVE improved discrimination consistently across 21 clinical tasks, with a mean ΔAUC of +0.090 (95% CI: 0.078-0.100; mean AUC 0.831 vs 0.742) over the Foundation VE, and 20 of 21 tasks improving with 95% CIs excluding zero (**Central Illustration**). Gains were largest for tasks the Foundation VE handled poorly, such as hypertrophic cardiomyopathy, and were not confined to 1 task type.

The VE approach can lower the barrier to adapting imaging AI locally but does not eliminate the need for the foundation model: producing the stored vectors requires the EchoPrime encoder, a shared cost. What it removes is the repeat cost. MVE training was completed in minutes on 1 workstation GPU, and once a study is stored as a single 512-dimensional vector, serving it needs neither the multigigabyte model weights nor the DICOM files the base case reads. The contribution is thus efficient downstream adaptation and storage rather than freedom from large-scale pretraining.[10,11]

The improvement was present at every level of examination completeness, including studies with fewer than 20 videos (**Figure 4C**). This matters clinically because real-world echocardiographic examinations vary in completeness because of patient factors, time constraints, or equipment limitations, so a study-level representation is useful in outpatient screening and emergency settings where a comprehensive examination may not be feasible.

TECHNICAL INNOVATIONS AND METHODOLOGICAL INSIGHTS. The consistent performance improvements achieved by the MVE VEs demonstrate the

CENTRAL ILLUSTRATION Multiview Integration Framework for Robust Echocardiographic Embedding



Study Population

- MIMIC-IV-ECHO: one academic center (Boston, MA, USA), 2017-2019
- 7,169 transthoracic studies from 4,555 patients
- Median 41 videos per study (IQR 34-48)
- 21 binary tasks: valvular disease, chamber size, ventricular function, hemodynamics, and ICD-coded diagnoses

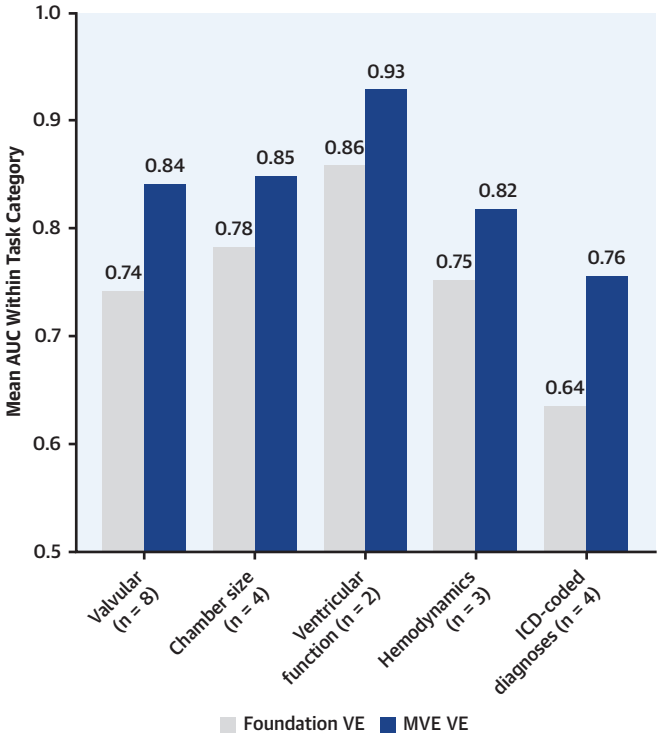
Multiview Encoder Framework

- Learned integration of all video embeddings into a single compact study-level representation (512 dimensions, 15-fold smaller than the Foundation VE)
- Small encoder (7.9 M parameters) trained on top of a pretrained foundation model's embeddings

Key Findings

- Mean AUC across 21 tasks: 0.831 vs 0.742 (difference 9.0 points, 95% CI: 7.8-10.0), higher in every task category
- Improvement retained in studies with <20 videos
- Adversarial learning reduced demographic predictability only at strong weights and at a substantial cost to clinical performance

Discrimination by Task Category



Conclusion: Integrating all videos of a study into one compact embedding improved discrimination across 21 tasks at minimal added cost; external validation is needed before clinical use.

Tohyama T, et al. JACC Adv. 2026;5(12):103488.

Echocardiographic videos are converted into video-level embeddings using a foundation model. A multiview encoder integrates these embeddings into study-level representations. These aggregated embeddings provide a compact yet comprehensive representation of cardiac function. The study population is summarized on the left. The bar chart shows the mean AUC within each task category for the Foundation VE and for the multiview encoder embedding without adversarial learning (point estimates; per-task values with 95% CIs are given in Supplemental Table 3). AI = artificial intelligence; MVE = multiview encoder; VE = vector embedding; ICD = International Classification of Diseases; MIMIC = Medical Information Mart for Intensive Care.

effectiveness of multiview integration using a VE approach. Unlike weighted integration based on classified views, which results in high-dimensional representations (15 × 512 dimensions per study),[3] MVE enables dynamic weighting of echocardiographic views through attention pooling mechanisms while compressing information into a single 512-dimensional study-level representation.

Importantly, MVE's attention mechanism enables more fine-grained adaptive integration without any explicit view classification. This aspect may be particularly effective for Doppler echocardiography and nonroutine views. Even within the same view category, color Doppler focusing on different blood flow regions and zoom-level variations cause substantial changes in the area of focus. Such variations cannot be captured by conventional approaches based on view classification; however, the MVE approach can finely adjust the importance based on the specific content and diagnostic value of each video.

This aggregation approach aligns with established clinical practice in which experienced cardiologists routinely integrate information from multiple echocardiographic views to achieve accurate diagnoses.[1] MVE attention weights can provide interpretability by revealing the views that contribute the most to individual predictions. However, as with other deep learning approaches, the model's internal decision-making process remains largely opaque because of complex nonlinear transformations in the transformer layers. Future work should incorporate explainability methods to enhance clinical trust.

DEMOGRAPHIC PREDICTABILITY AND ADVERSARIAL LEARNING. Adversarial learning did not remove demographic information at an acceptable cost. Varying the adversarial weight from 0 to 1.0 reduced demographic predictability only at weights that also degraded clinical discrimination, and the one marked task-level gap narrowed only partially (**Figure 3**). These results indicate that gradient reversal, as implemented here, did not achieve a favorable trade-off on this data set. Class imbalance across race categories, systematic differences in body habitus and in the views acquired, label bias in a single clinical read, and residual confounding by comorbidity, indication, and care setting may all contribute, and our design cannot distinguish among them. What our design does support is that the MVE did not add demographic information or widen subgroup differences: demographic predictability was similar to that of the Foundation VE (**Figure 2B**), and the MVE advantage was uniform across sex, race, and video-count subgroups, with every difference-in-differences interval spanning zero (**Figure 4**). Demographic information and subgroup differences therefore appear to originate in the upstream representation.[26]

CLINICAL IMPLEMENTATION AND PRACTICAL CONSIDERATIONS. For clinical translation, MVE requires users to have a basic understanding of machine learning workflows and VE concepts, combined with clinical expertise, to assess echocardiographic video quality and interpret model outputs within an appropriate clinical context. This aligns with current expectations for AI-assisted cardiology tools,[28,29] which require clinical expertise and basic AI literacy, but not specialized machine learning knowledge.

Several data-quality considerations must be addressed to ensure successful implementation. Poor-quality videos with excessive noise, inappropriate gain settings, or suboptimal imaging windows may result in unreliable VEs and should be flagged for manual review.[1,28]

Adapting the model downstream requires only standard research computing rather than industrial-scale infrastructure. An institution that can obtain a pretrained encoder, or embeddings produced with it, could therefore develop its own models without large computing capacity.[10,11] Developing methods to extract and refine representations from existing models may be a more accessible path to advanced capabilities than continuously scaling model sizes.

STUDY LIMITATIONS. This study had several important limitations. First, and most importantly, this was a single-center retrospective study without external validation and therefore cannot support clinical adoption. All studies come from one academic center over a restricted period (2017-2019), with a limited range of vendor configurations, acquisition protocols, and reporting conventions, so the results are a methodological demonstration on this data set rather than evidence of clinical readiness. Although open-source echocardiography data sets are available, most are limited in size or restricted to specific imaging views. Future efforts toward the public release of multiview echocardiographic data sets by more institutions and the implementation of multi-institutional validation studies are essential for establishing broader generalizability. Second, our framework inherits EchoPrime's limitations and biases. Because the MVE starts from the video embeddings and never sees the pixels, any bias the foundation model encodes is carried into

our representation, and information the foundation model fails to capture for a given view cannot be recovered downstream. Third, the reference standard is imperfect and varies in quality across tasks. Measurement-derived labels come from routine clinical interpretation rather than standardized research reread, and the databases do not record whether more than one reader contributed. Missingness varied widely and is driven by clinical indication, so it is unlikely to be completely at random, and each task is therefore evaluated in the subpopulation in which that measurement or diagnosis was clinically pursued. Fourth, our study focused on gradient reversal layer-based adversarial learning.[25] Other approaches such as disentangled representation learning or data reweighting[30-32] may be more effective in mitigating demographic biases and warrant future investigation.

CONCLUSIONS

Our MVE framework integrates multiple echocardiographic views into unified study-level embeddings, enabling substantial diagnostic improvements across diverse cardiac tasks. By producing compact, informative embeddings without requiring extensive computational resources, this approach lowers the cost of building echocardiographic AI on an existing foundation model.

DATA AVAILABILITY STATEMENT. The MIMIC-IV and MIMIC-IV-ECHO data sets analyzed in this study are publicly available at PhysioNet (https://physionet.org/), including the structured echocardiographic measurements used to define the outcomes, which are distributed in MIMIC-IV-ECHO version 1.0. The data sets used in this study cannot be directly redistributed because of patient privacy protection but can be obtained through the established PhysioNet access procedures.

CODE AVAILABILITY STATEMENT. Source code for MVE implementation, training procedures, and evaluation scripts is publicly available at https://github.com/ttohya/mve-echo under the MIT License.

ACKNOWLEDGMENTS The authors thank the MIMIC-IV and MIMIC-IV-ECHO data contributors and the PhysioNet team for making these data sets publicly available. Large language models were used to improve the readability and grammar of the manuscript. All intellectual contributions, data analyses, and scientific interpretations were performed by the authors.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

Dr Tohyama was supported by a grant from the Japan Foundation for Applied Enzymology and the Uehara Memorial Foundation Fellowship. Dr Celi is funded by the National Institutes of Health through DS-I Africa U54 TW012043-01 and Bridge2AI OT2OD032701, the National Science Foundation through ITEST #2148451, and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2024-00403047). The funders had no role in study design, data collection, analysis, interpretation of data, or writing of this manuscript. Dr Tohyama joined MeDiCU Inc as an employee after the initial submission of this paper. The company is unrelated to the work reported here: it provided no funding, materials, or input, and had no role in the study or in the decision to submit it for publication. All other authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Jacques Kpodonu, Division of Cardiac Surgery, Beth Israel Deaconess Medical Center, 110 Francis Street, Boston, Massachusetts 02215, USA. E-mail: jkpodonu@bidmc.harvard.edu. X handle: @taketohyama.

PERSPECTIVES

COMPETENCY IN SYSTEMS-BASED PRACTICE: Reusing a single compact representation of an echocardiographic study for many diagnostic questions may allow centers without large computing resources to develop and maintain their own interpretation models, including for examinations with fewer videos.

TRANSLATIONAL OUTLOOK: Multi-institutional external validation and prospective evaluation across ultrasound vendors and patient populations are needed before such locally developed models are used clinically, together with methods that reduce demographic information in the embedding without degrading diagnostic performance.

REFERENCES

1. Mitchell C, Rahko PS, Blauwet LA, et al. Guidelines for performing a comprehensive transthoracic echocardiographic examination in adults: recommendations from the American Society of echocardiography. *J Am Soc Echocardiogr*. 2019;32:1-64.

2. Holste G, Oikonomou EK, Tokodi M, et al. Complete AI-Enabled echocardiography interpretation with multitask deep learning. *JAMA*. 2025;334:306-318.

3. Vukadinovic M, Chiu IM, Tang X, et al. Comprehensive echocardiogram evaluation with view primed vision language AI. *Nature*. 2026;650:970-977.

4. Chung DJ, Lee SM, Kaker V, et al. Echocardiogram vector embeddings via R3D transformer for the advancement of automated echocardiography. *JACC Adv*. 2024;3:101196.

5. Christensen M, Vukadinovic M, Yuan N, et al. Vision-language foundation model for echocardiogram interpretation. *Nat Med*. 2024;30:1481-1488.

6. Rajpurkar P, Chen E, Banerjee O, et al. AI in health and medicine. *Nat Med*. 2022;28:31-38.

7. Strubell E, Ganesh A, Mccallum A. *Energy and policy Considerations for Deep Learning in NLP*. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019: 3645-3650.

8. Zhou J, Du M, Chang S, et al. Artificial intelligence in echocardiography: detection, functional evaluation, and disease diagnosis. *Cardiovasc Ultrasound*. 2021;19:1-11.

9. Nabi W, Bansal A, Xu B. Applications of artificial intelligence and machine learning approaches in echocardiography. *Echocardiography*. 2021;38: 982-992.

10. Moukheiber D, Restrepo D, Cajas SA, et al. A multimodal framework for extraction and fusion of satellite images and public health data. *Sci Data*. 2024;11:1-20.

11. Sellergren AB, Chen C, Nabulsi Z, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*. 2022;305: 454-465.

12. Yang Y, Zhang H, Gichoya JW, et al. The limits of fair medical imaging AI in real-world generalization. *Nat Med*. 2024;30:2838-2848.

13. Yang J, Soltan AAS, Eyre DW, et al. An adversarial training framework for mitigating algorithmic biases in clinical machine learning. *NPJ Digit Med*. 2023;6:1-10.

14. Ricci Lara MA, Echeveste R, Ferrante E. Addressing fairness in artificial intelligence for medical imaging. *Nat Commun*. 2022;13: 1-6.

15. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10:1-9.

16. Johnson A, Bulgarelli L, Pollard T, et al. MIMIC-IV (version 3.1). *PhysioNet*. 2024. https://doi.org/10.13026/kpb9-mt58

17. Gow BJ, Han A, Greenbaum NR, et al. MIMIC-IV-Echo: an echocardiography dataset integrated with the MIMIC-IV ecosystem. *Sci Data*. 2026. https://doi.org/10.1038/s41597-026-08386-7

18. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*. 2024;385:e078378. https://doi.org/10.1136/BMJ-2023-078378

19. Sengupta PP, Shrestha S, Berthon B, et al. Proposed requirements for cardiovascular imaging-related machine learning evaluation (PRIME): a checklist: reviewed by the American College of Cardiology Healthcare Innovation Council. *JACC Cardiovasc Imaging*. 2020;13:2017-2035.

20. Lang RM, Badano LP, Victor MA, 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. *J Am Soc Echocardiogr*. 2015;28:1-39.e14.

21. Nagueh SF, Smiseth OA, Appleton CP, et al. Recommendations for the evaluation of left ventricular diastolic function by echocardiography: an update from the American Society of echocardiography and the European Association of cardiovascular imaging. *J Am Soc Echocardiogr*. 2016;29:277-314.

22. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:5998-6008.

23. Sanh V, Debut L, Chaumond J, et al. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. Accessed July 14, 2025. https://arxiv.org/pdf/1910.01108

24. Sun Z, Yu H, Song X, et al. *MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices*. Proceedings of the Annual Meeting of the Association for Computational Linguistics; 2020:2158-2170.

25. Ganin Y, Ustinova E, Ajakan H, et al. Domain-Adversarial training of neural networks. *J Machine Learn Res*. 2016;17:1-35.

26. Brown A, Tomasev N, Freyberg J, et al. Detecting shortcut learning for fair medical AI using shortcut testing. *Nat Commun*. 2023;14:1-10.

27. Zhang BH, Lemoine B, Mitchell M. *Mitigating Unwanted Biases with Adversarial Learning*. AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society; 2018:335-340.

28. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25:44-56.

29. Ouyang D, He B, Ghorbani A, et al. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*. 2020;580:252-256.

30. Yang Y, Lin M, Zhao H, et al. A survey of recent methods for addressing AI fairness and bias in biomedicine. *J Biomed Inform*. 2024;154: 104646.

31. Liu X, Sanchez P, Thermos S, et al. Learning disentangled representations in the imaging domain. *Med Image Anal*. 2022;80:102516.

32. Krasanakis E, Spyromitros-Xioufis E, Papadopoulos S, et al. *Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-Aware Classification*. The Web Conference 2018 - Proceedings of the World Wide Web Conference, WWW 2018; 2018:853-862.

KEY WORDS artificial intelligence, computational efficiency, echocardiography, shortcut learning, vector representation

APPENDIX For an expanded Methods section as well as supplemental figures and tables, please see the online version of this paper.