EchoStature: Spatial Demographic Conditioning for Pediatric Ejection Fraction Regression, with Subgroup-Stratified Evaluation

Siddhi Narayan[†1], Khadeeja Hussain[†1], Sai Phani Parsa[2], Jahnavi Veeramachinani[3], Mohammad Husain[*1], Sai Chandra Kosaraju[*1]

[1]*Department of Computer Science, California State Polytechnic University, Pomona, California, USA*

[2]*Department of Computer Science, University of Nevada Las Vegas, Las Vegas, Nevada, USA*

[3]*Department of Information Systems, University of Nevada Las Vegas, Las Vegas, Nevada, USA*

[*]*Corresponding Authors: mihusain@cpp.edu, skosaraju@cpp.edu*

Left ventricular ejection fraction (EF) is the principal measure of systolic function in pediatric echocardiography, and deep video models now estimate it at accuracy comparable to expert readers. Pediatric cohorts nonetheless span an order of magnitude in body size, and reported accuracy is pooled across that range. Body size, rather than age or sex, governs pediatric cardiac dimensions, and it acts on echocardiograms through geometry rather than as a scalar offset, yet demographic conditioning in EF regression conventionally encodes it as a low-dimensional vector combined after spatial pooling.

We present EchoStature, which uses gated fusion to integrate real echocardiograms with demographically conditioned 3D reconstructions while limiting reconstruction artifacts. A hierarchical contrastive objective further organizes the latent space by EF and demographic subgroup. On EchoNet-Pediatric, EchoStature matches the video-only reference model in overall MAE (4.49) and remains stable when demographic attributes are ablated at inference. However, subgroup analysis reveals substantially higher error in infants than in children aged 2-5 years (5.79 vs. 3.89 MAE), highlighting age-dependent performance disparities that are obscured by pooled metrics. Pooled accuracy therefore conceals clinically meaningful variation, particularly among the youngest patients, for whom echocardiography is often the only available assessment of ventricular function. Accordingly, we report subgroup-stratified error for every configuration and argue that such reporting should be standard in pediatric EF regression. Code availability: `https://github.com/KosarajuBioAI/EchoStature`

Keywords: Pediatric echocardiography; Ejection fraction regression; Subgroup performance; Demographic conditioning; Gated feature fusion; Conditional video generation; Counterfactual analysis; Explainable AI.

[†]Authors contributed equally to this work.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.This work is partially supported by Mohammad Husain and Sai Chandra Kosarajus Cal Poly Pomona Internal Grant

1. Introduction

Left ventricular ejection fraction (EF) anchors systolic assessment, guiding cardiomyopathy diagnosis, heart-failure staging, and therapy timing. Echocardiography, the frontline cardiac test in children, is the primary route to EF because it is fast, portable, and free of ionizing radiation,[1,2] but manual estimates are imprecise and vary substantially between readers.[1] Pediatric cardiac structure changes rapidly with age, and EchoNet-Pediatric spans this range with 4,467 studies from 1,958 patients aged 0–18 years.[2]

EchoNet-Dynamic reached expert-level accuracy (mean absolute error [MAE] 4.1%) by regressing EF with factorized spatiotemporal convolutions,[1,3] and later models refined frame selection,[4] anatomical grounding,[5,6] and temporal aggregation.[7] Pediatric retraining outperformed an adult model ($P < .001$; MAE 3.66% with two views), yet pooled accuracy hides subgroup variation: 86% of EchoNet-Pediatric studies have EF ≥55%, and sex, age, and body habitus are unevenly represented.[2] Imbalanced training degrades performance on underrepresented groups in classification and cardiac segmentation,[8,9] and chest radiograph classifiers reportedly underdiagnose under-served subgroups,[10] though this was later contested. Pediatric EF regression therefore demands subgroup-stratified reporting.

Body size governs pediatric cardiac dimensions more strongly than age or sex,[11] and it acts spatially, altering chamber size, sector geometry, and the apparent scale of ventricular motion. Existing fusion does not preserve this: demographic vectors join video features only after pooling has discarded spatial structure,[12] and broadcast vectors stay spatially uniform. In semantic synthesis, such uniform conditioning washes out with depth, unlike spatially varying conditioning.[13] Feature-wise modulation is a partial remedy but also acts uniformly across positions,[14] and short metadata-to-label routes invite shortcut learning.[15]

Conditional generation can instead express attributes as image structure,[16] in video through three-dimensional convolutions[17] and in echocardiography through D'ARTAGNAN, which alters EF while preserving patient anatomy and imaging style.[18] Generated data, however, mislead when realism substitutes for utility,[19] and generator artifacts inject spurious correlations,[20] so their influence must be limited. Contrastive objectives structure imbalanced latent spaces,[21,22] and Grad-CAM and counterfactual probes test whether predictions rely on cardiac evidence rather than metadata shortcuts.[23–26]

These threads converge on design requirements: demographics should reach the encoder as spatial structure rather than a scalar offset, any generated stream should stay subordinate to the real acquisition, and the latent space should be organized by both EF and demographic subgroup. Evaluation should pair parameter-sensitive attribution with controlled demographic perturbation and report error by subgroup.

2. Methodology

2.1. *Overview*

EchoStature estimates pediatric EF from an echocardiographic clip, its demographic attributes, and a demographically rendered synthetic clip. The real clip and the synthetic rendering pass through a shared frame encoder, and a learned gate combines their frame-level

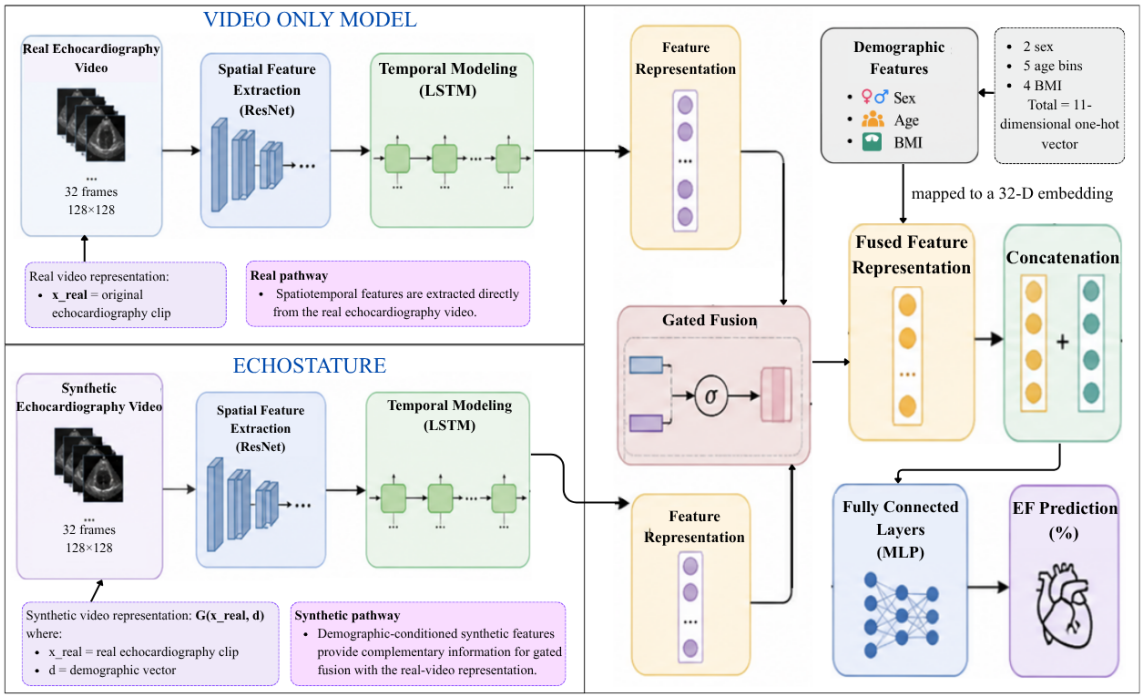


Fig. 1: Overview of the proposed ECHOSTATURE model. The benchmark model estimates ejection fraction (EF) from a real echocardiography video and demographic embedding, whereas ECHOSTATURE additionally incorporates a paired synthetic reconstruction through gated feature fusion to supply a rendered demographic stream.

embeddings before temporal aggregation. Demographic attributes additionally enter as a low-dimensional embedding concatenated with the pooled video representation, following standard practice in multimodal EF regression.[12] Weight sharing across the two streams keeps them in a common embedding space and confines the additional parameters to the gating network. Figure 1 summarizes the architecture.

2.2. *Feature representation*

Each echocardiographic clip is resized to 128×128 pixels, uniformly sampled or zero-padded to a fixed length of $T = 32$ frames, and normalized to $[0, 1]$. Demographic attributes supply contextual information that the clip does not directly reveal, since acquisition geometry and chamber scale depend on body size.[11] Sex, age bin, and BMI category are encoded as an 11-dimensional one-hot vector comprising two sex categories, five age bins, and four BMI categories,

$$d \in \{0, 1\}^{11}, \tag{1}$$

$$e = \psi(d) \in \mathbb{R}^{32}, \tag{2}$$

Here, $\mathbf{d}$ denotes the demographic attribute vector, $\psi(\cdot)$ denotes a lightweight multilayer perceptron, and $\mathbf{e}$ denotes the resulting 32-dimensional demographic embedding. The same de-

mographic vector conditions the C3DGAN generator, which produces a paired reconstruction of the real echocardiographic clip, as described below.

C3DGAN. The synthetic clip used by EchoStature is produced by a conditional 3D reconstruction GAN (C3DGAN). The generator receives the real echocardiographic clip x_{real} together with its 11-dimensional demographic vector d, and produces a paired reconstruction

$$x_{\text{syn}} = G(x_{\text{real}}, d). \tag{3}$$

The generator therefore reconstructs the observed clip under its matched demographic condition rather than sampling a clip from latent noise. It operates on grayscale clips of 16 frames at 64×64 and preserves the temporal axis throughout. After generation, each stored reconstruction is resampled with its paired real clip to $T = 32$ frames at 128×128 for EF training. The checkpoint used here corresponds to epoch 183. EchoStature does not call G during EF training: it loads the precomputed paired clip and inherits the EF label of x_{real}.

Generator. G is a 3D U-Net with four encoder levels of 64, 128, 256, and 512 channels. The demographic vector is mapped by a small MLP to a 128-dimensional embedding and fused into the video features by projection, broadcast, and $1 \times 1 \times 1$ convolution at the first three encoder levels and at the bottleneck. Spatial downsampling is used in the encoder while the temporal length is kept fixed. Residual blocks use $3 \times 3 \times 3$ convolutions with squeeze-and-excitation channel attention. The decoder uses transposed convolutions, U-Net skips, and a final tanh to emit a single-channel clip. The saved generator contains approximately 1.07×10^8 parameters.

Discriminator. The discriminator is a spectral-normalized 3D PatchGAN that scores real versus reconstructed video and does not take d as a direct input. Four spatial downsampling stages use 64, 128, 256, and 512 channels; a convolutional head produces patch-level real/fake scores, and an auxiliary head maps a pooled 512-dimensional representation to an 11-way demographic prediction. The saved discriminator contains approximately 2.90×10^6 parameters.

Training and use in EchoStature. Generator and discriminator were trained with Adam at learning rate 10^{-4} and $\beta = (0.5, 0.999)$. The checkpoint stores a single combined loss value and does not retain the individual loss terms or their weights, so those coefficients are not reported here. After training, one reconstruction was written for each real clip used by EchoStature; the EF model therefore sees a fixed paired rendered stream rather than an on-the-fly sample from G. This stream is the spatiotemporal demographic pathway that the gate in Section 2.3 can admit or suppress; it is complementary to the post-pooling embedding $e = \psi(d)$, which remains a low-dimensional scalar pathway.

2.3. *Architecture and training*

Let $\phi(\cdot)$ denote the frame encoder, comprising a 2D convolutional backbone followed by a linear projection to dimension d_f which is the frame-embedding dimension. The same encoder is applied to corresponding frames of both streams,

$$r_t = \phi\Big(x_t^{\text{real}}\Big) \in \mathbb{R}^{d_f}, \tag{4}$$

$$\tilde{r}_t = \phi(x_t^{\text{syn}}) \in \mathbb{R}^{d_f}, \tag{5}$$

for $t = 1, \ldots, T$. Frame-wise encoding admits a gate that operates at every time step, allowing the model to vary its reliance on the rendered stream across the cardiac cycle.

A gating network $\gamma(\cdot)$ estimates the per-dimension contribution of each stream at every frame,

$$c_t = [r_t; \tilde{r}_t] \in \mathbb{R}^{2d_f}, \tag{6}$$

$$g_t = \sigma(\gamma(c_t)) \in [0,1]^{d_f}, \tag{7}$$

$$f_t = g_t \odot r_t + (1 - g_t) \odot \tilde{r}_t, \tag{8}$$

where $\odot$ denotes element-wise multiplication and $\sigma(\cdot)$ denotes the element-wise sigmoid. Here, $\mathbf{r}_t$ and $\tilde{\mathbf{r}}_t$ denote the frame-level representations of the real and synthetic streams, respectively, $\mathbf{c}_t$ is their concatenated representation, $\mathbf{g}_t$ is the learned gate, and $\mathbf{f}_t$ is the resulting fused representation at time step t. The convex form bounds the synthetic contribution and allows the network to suppress the rendered stream entirely as $g_t \to 1$, which limits the propagation of reconstruction artifacts.[19]

A recurrent temporal encoder $\tau(\cdot)$, instantiated as a bidirectional long short-term memory network with hidden dimension d_h per direction, models the dynamics of the fused sequence. Temporal attention then aggregates the resulting states into a single video representation,

$$z_t = \tau\big(f_t \mid \{f_k\}_{k=1}^{T}\big) \in \mathbb{R}^{2d_h}, \tag{9}$$

$$a_t = \frac{\exp(w^\top z_t)}{\sum_{k=1}^{T} \exp(w^\top z_k)}, \tag{10}$$

$$p = \sum_{t=1}^{T} a_t z_t, \tag{11}$$

where $w \in \mathbb{R}^{2d_h}$ is a learnable attention vector shared across time steps, $\mathbf{z}_t$ is the bidirectional temporal state at frame t, a_t is its attention weight, and $\mathbf{p}$ is the temporally pooled video representation. The pooled representation is concatenated with the demographic embedding, $u = [p; e]$, and passed through a regression head $\rho(\cdot)$ with sigmoid activation,

$$\hat{y} = \sigma(\rho(u)) \in [0,1], \qquad \widehat{\text{EF}} = 100\,\hat{y}. \tag{12}$$

Here, y is the ground-truth EF divided by 100, so $y \in [0,1]$, and $\hat{y}$ is the predicted value on that same scale. $\widehat{\text{EF}} = 100\,\hat{y}$ is the prediction in percentage points. Training combines supervised regression with a hierarchical contrastive objective. The regression loss balances robustness against squared-error optimization,

$$\mathcal{L}_{\text{EF}} = (1 - w_{\text{r}})\,\mathcal{L}_\delta(\hat{y}, y) + w_{\text{r}}\,\|\hat{y} - y\|_2^2, \tag{13}$$

where $\mathcal{L}_\delta$ denotes the Huber loss and $w_{\text{r}} \in [0,1]$ sets the balance between the two terms. The contrastive objective operates on the projected video embedding and proceeds in two

Table 1: Summary statistics of the EchoNet-Pediatric cohort.

Characteristic	**Train**	**Validation**
Number of clips	6,227	1,558
EF (%, mean ± SD)	61.1 ± 10.5	61.4 ± 9.7
Sex (F/M/O)	2661 / 3561 / 5	653 / 905 / 0
View (A4C/PSAX)	2657 / 3570	613 / 945

F: female; M: male; O: other; A4C: apical four-chamber; PSAX: parasternal short-axis.

stages. The first stage, $\mathcal{L}_{\mathrm{H1}}$, treats samples sharing an EF-derived task label as positives, following supervised contrastive learning.[22] The second stage, $\mathcal{L}_{\mathrm{H2}}$, refines the representation by treating samples sharing a demographic subgroup as positives within each EF group, which counteracts the concentration of pediatric labels in the normal range.[2] The combined auxiliary loss is

$$\mathcal{L}_{\mathrm{HCL}} = \mathcal{L}_{\mathrm{H1}} + \lambda_{\mathrm{H2}}(\epsilon)\, \mathcal{L}_{\mathrm{H2}}, \tag{14}$$

where $\lambda_{\mathrm{H2}}(\epsilon)$ increases with epoch ϵ, so that EF structure forms before subgroup structure is imposed. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{EF}} + \alpha\, \mathcal{L}_{\mathrm{HCL}}, \tag{15}$$

where α controls the contribution of the hierarchical contrastive objective, L_{EF} is the supervised regression loss, L_{δ} is the Huber loss, L_{H1} is the EF-based supervised contrastive loss, and L_{H2} is the demographic hierarchical contrastive loss. The parameter λ_{H2} controls the contribution of the second contrastive stage and varies with training epoch ϵ. The model is optimized with AdamW under cosine learning-rate scheduling.

2.4. *Benchmark and implementation*

We evaluate on EchoNet-Pediatric, which comprises 7,643 apical four-chamber and parasternal short-axis clips drawn from 4,467 studies across 1,958 patients aged 0 to 18 years.[2] As a benchmark reference we train a video-only regression model that shares the frame encoder, temporal encoder, attention module, and regression head of EchoStature. The two models do not receive the same input. The reference model observes the echocardiographic clip alone, whereas EchoStature additionally observes the demographic vector and the synthetic clip rendered from it. The comparison therefore evaluates supplying demographic context in both scalar and rendered form against a video-only reference; it does not isolate either form individually, and under the multi-seed protocol reported in Section 3.2 it does not establish a pooled accuracy gain. Both models use AdamW, learning rate 3×10^{-4}, a cosine schedule, batch size 8, the same number of epochs, and the same clip-level split. The contrastive weight is 0.04 for the video-only model and 0.025 for EchoStature, and EchoStature adds the gating network, so this comparison is not a single-component ablation.

3. Experiments and results

3.1. *Dataset and experimental setup*

We evaluate on EchoNet-Pediatric, predicting left ventricular ejection fraction from pediatric echocardiography videos.[2] Each sample comprises a processed grayscale clip acquired in the apical four-chamber (A4C) or parasternal short-axis (PSAX) view, together with the corresponding EF, sex, age band (0–1, 2–5, 6–10, 11–15, and 16–18 years), and BMI category. BMI categories are derived from fixed body-mass-index thresholds rather than age- and sex-specific pediatric percentiles, since percentile references are undefined below two years of age and the cohort spans patients from birth. The resulting categories stratify patients by raw body mass index and do not correspond to clinical weight-status definitions. After preprocessing and quality filtering, the cohort contains 7,785 labeled clips, partitioned at the clip level into 6,227 training and 1,558 validation clips. The validation set is an internal validation set rather than an external cohort. Because the partition is performed at the clip level, multiple clips from the same patient may potentially occur across the training and validation sets. Therefore, the reported validation results should not be interpreted as evidence of patient-disjoint generalization. Table 1 summarizes the partitions, and all reported results are obtained on the validation set. Each clip is resized to 128×128 pixels, uniformly sampled to $T = 32$ frames, and normalized to $[0, 1]$. EchoStature additionally receives one demographically rendered synthetic clip for each real clip. The synthetic clip is paired with its real counterpart and shares the counterpart's EF label and demographic attributes; it is therefore treated as demographic-conditioned visual context rather than as an independently labeled clinical observation, while the video-only reference model observes the echocardiographic clip alone. Each paired clip was compared with its source clip by structural similarity and peak signal-to-noise ratio. Across the 7,785 clips used for training and validation, mean structural similarity is 0.995 and mean peak signal-to-noise ratio is 49.0 dB. This comparison measures agreement with the source when the generator receives that source and its own demographic vector. It does not test whether a different demographic vector would change the clip, and it does not establish clinical realism or clinical validity.

3.2. *Overall performance*

Table 2 reports regression performance on the validation set. Under seven matched 100-epoch seeds, EchoStature reaches 4.49 ± 0.04 MAE, 6.17 ± 0.07 RMSE, and $R^2 = 0.587 \pm 0.010$, compared with 4.49 ± 0.05, 6.17 ± 0.06, and 0.588 ± 0.008 for the video-only reference. The pooled difference is therefore negligible under this protocol. The subgroup analysis that follows shows that age-group variation remains far larger than this pooled architecture gap.

3.3. *Subgroup performance*

Table 4 reports MAE (mean ± SD over seven seeds) across sex, age, and body-mass subgroups, and the variation it reveals exceeds every architectural effect measured in this work. Infants aged 0-1 years reach 5.79 ± 0.20 MAE against 3.89 ± 0.14 for the 2-5 year group. That difference of about 1.9 percentage points is far larger than the negligible pooled margin between

EchoStature and the video-only reference. Where a patient falls in the pediatric age range therefore affects expected error far more than which model estimates it.

The observed performance gap may reflect characteristics of infant echocardiography, including differences in heart rate, chamber size, and the amount of anatomical information represented at the fixed spatial and temporal resolution used in this study. For example, higher infant heart rates could cause a fixed 32-frame sample to represent a different portion of the cardiac cycle than in older patients, while smaller chamber dimensions could occupy fewer pixels at the fixed input resolution. These factors are plausible explanations for the observed subgroup difference, but their individual contributions were not directly tested in the present analysis.

Differences across sex and body-mass strata are smaller than the age gap. Female patients reach 4.23 ± 0.07 MAE against 4.68 ± 0.07 for male patients, and the underweight stratum reaches 4.38 ± 0.11 against 5.17 ± 0.15 for overweight. Subgroup counts are unequal, so per-subgroup estimates carry correspondingly different precision, and the smaller strata warrant cautious interpretation.

The practical implication is that pooled accuracy misrepresents pediatric performance. A model reported near 4.49 MAE overall still performs near 5.8 MAE for the youngest patients, who are also those for whom echocardiography is most often the only available assessment of ventricular function.[2]

Reporting a single pooled figure conceals this variation entirely, and it conceals it for the subgroup least able to tolerate error. Pediatric EF regression therefore warrants subgroup-stratified reporting as standard practice, independently of the architecture under evaluation.

3.4. *Ablation studies*

Gated fusion. Table 3 compares variants trained under a matched 100-epoch budget without the contrastive objective. The three rows share that 100-epoch budget, AdamW, learning rate 3×10^{-4}, a cosine schedule, batch size 8, the same clip-level split, and a contrastive weight of zero, and are reported as mean ± SD over five seeds. The real-only and rendered-only rows use the same video-only network and differ in which clips are supplied. The real-plus-rendered row uses the EchoStature network, including the gate. Under five matched 100-epoch seeds without the contrastive objective (Table 3), real-only reaches 4.44 ± 0.07 MAE, real-plus-rendered reaches 4.46 ± 0.15, and rendered-only reaches 4.66 ± 0.03. The gated stream therefore does not improve pooled MAE over real-only under this protocol, while rendered-only remains worse than real video, indicating that the rendered clips retain EF-relevant structure but are not a substitute for real acquisitions.

Contrastive objective. Under the same 100-epoch budget, the five-seed no-contrastive real-plus-rendered model reaches 4.46 ± 0.15 MAE (Table 3), compared with 4.49 ± 0.04 for the seven-seed full EchoStature configuration (Table 2). Because the seed counts differ and several components change together, these figures do not isolate the contrastive term. The objective remains part of the full training configuration, and its contribution is not independently established.

Demographics-only control. Replacing the learned video representation entering the re-

gression head with zeros, while retaining the demographic embedding, degrades performance substantially: on a representative checkpoint from the seven-seed protocol in Table 2, MAE rises from about 4.49 to 7.61 for the video-only reference and from about 4.49 to 10.69 for EchoStature. This control does not retrain either model. At inference, the pooled video vector is set to zero and the demographic embedding is left unchanged. Both models therefore rely predominantly on cardiac video rather than on demographic metadata.

Demographic conditioning.

Table 5 reports inference-time perturbation of the demographic vector with the video inputs held fixed. Across seven seeds, zeroing the vector raises MAE from 4.49 ± 0.04 to 4.53 ± 0.04, and shuffling leaves MAE at 4.49 ± 0.04, so the change is at most about 0.04 MAE. This perturbation affects only the scalar demographic pathway, since the synthetic clip is rendered before inference and is unchanged by the manipulation, so the result bounds the contribution of the post-pooling embedding rather than that of the rendered stream.

3.5. *Rendered-stream chamber structure*

Pixel-level agreement alone does not establish a coherent beating cavity in the rendered stream. On 100 random validation pairs, a classical dark-cavity connected-component check detects a cavity in all 32 frames of every real and rendered clip. Relative area change is 0.65 ± 0.11 (real) vs. 0.61 ± 0.12 (rendered; Kolmogorov–Smirnov $p = 0.054$), with associated within-pair ranks (Spearman $\rho = 0.72$). Figure 2 summarizes the check. We treat this as image-analysis evidence of a detectable cavity under paired reconstruction, not as clinical left-ventricle segmentation.

3.6. *Model interpretation*

Representation structure. Figure 3b(ii) presents UMAP projections of the regressor input representation, colored by sex, age bin, and body-mass stratum. Partial separation is visible for sex and age bin. Body-mass strata remain largely intermixed in this projection even when points are colored by the true weight/height BMI categories, which is consistent with the coarse threshold-based encoding described in Section 3.1. We present these projections as a qualitative description rather than as evidence of a learned effect, since demographic attributes are partly inferable from echocardiograms directly[24] and since UMAP geometry depends on the chosen neighborhood parameters.

Attribution and counterfactual stability. Grad-CAM maps from both models concentrate on the imaging sector over cardiac structures rather than off-sector background.[25] Figure 3b shows predictions under controlled perturbation of sex, age bin, and BMI with the video held fixed. Predicted EF shifts only marginally, consistent with Table 5, while attribution maps can still change under the same perturbations.

4. Discussion

The dominant source of error variation in pediatric EF regression is the patient, not the architecture. The difference between the best and worst age subgroups in our experiments is far larger than the negligible pooled difference between the two models, and infants incur the

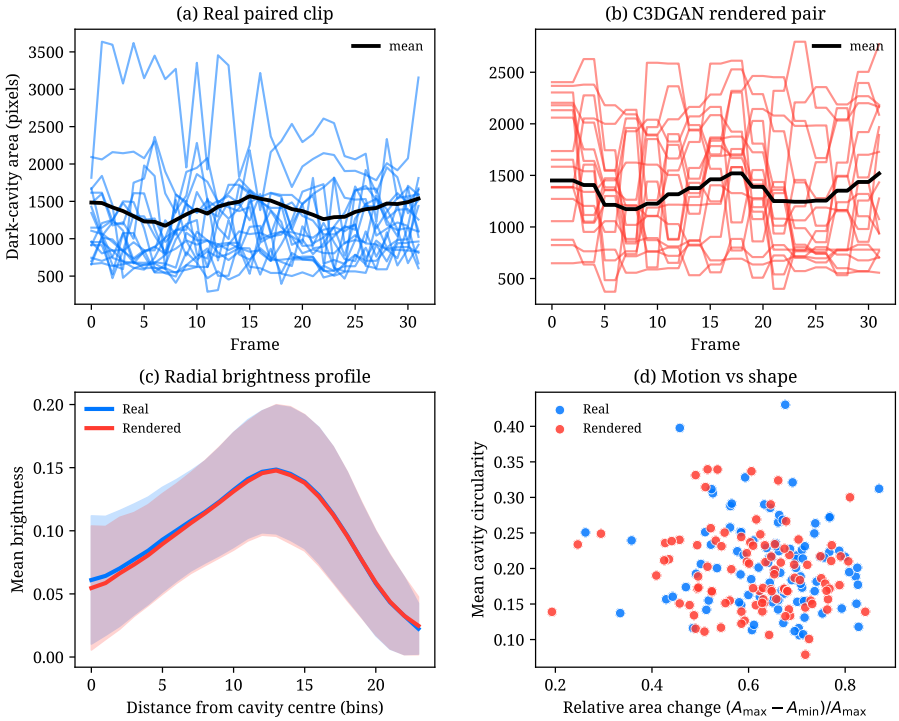


Fig. 2: Chamber structure in paired real and C3DGAN-rendered validation clips used by EchoStature. (a, b) Dark-cavity area over 32 frames for 100 real clips and their rendered pairs (thin curves: random subset; black: mean). (c) Radial brightness profile from the cavity centre. (d) Relative area change versus mean cavity circularity.

Table 2: Overall EF regression on the validation set (mean ± SD over 7 seeds, 100 epochs).

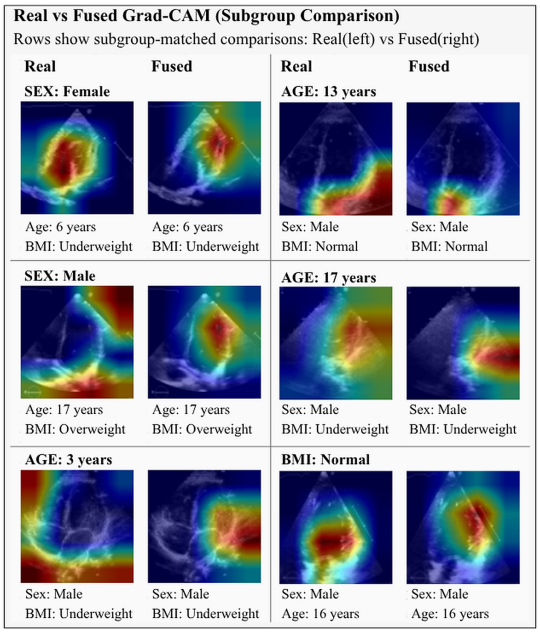
Model	MAE↓	RMSE↓	R^2 ↑
Video-only	4.49 ± 0.05	6.17 ± 0.06	0.588 ± 0.008
EchoStature	$\mathbf{4.49 \pm 0.04}$	$\mathbf{6.17 \pm 0.07}$	$\mathbf{0.587 \pm 0.010}$

Table 3: Input ablation under a matched 100-epoch schedule without the contrastive objective (mean ± SD over 5 seeds).

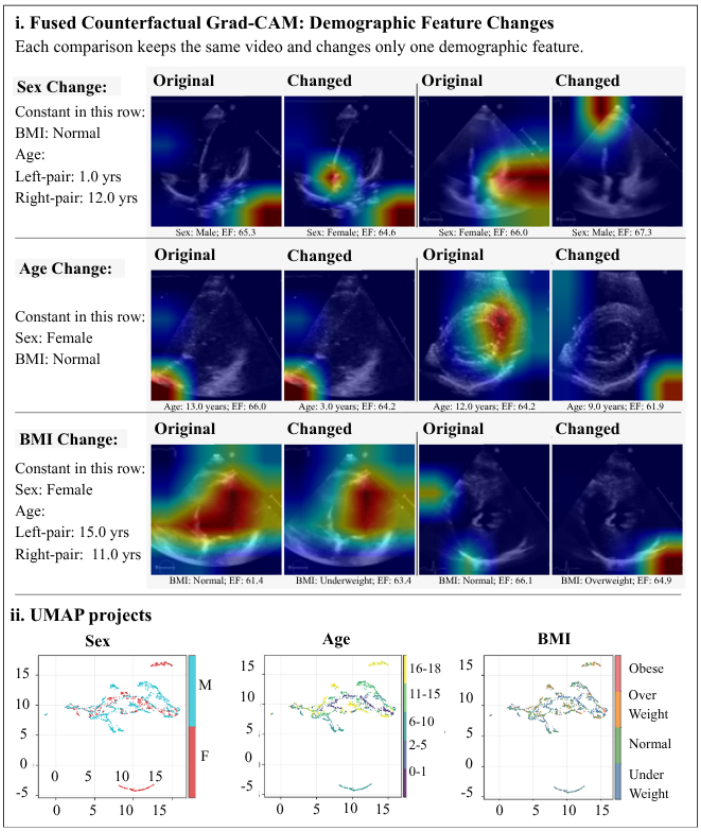
Input streams	Fusion	MAE↓	RMSE↓	R^2 ↑
Real + rendered	✓	4.46 ± 0.15	6.15 ± 0.28	0.590 ± 0.037
Real only	—	4.44 ± 0.07	6.07 ± 0.10	0.601 ± 0.013
Rendered only	—	4.66 ± 0.03	6.38 ± 0.07	0.564 ± 0.010

highest error of any subgroup. Architectural comparison on pooled metrics therefore optimizes a quantity that obscures the failure mode most relevant to deployment, since a model deployed across a pediatric service encounters the full age range and performs measurably worse at one end of it. We report subgroup-stratified error for every configuration in this work and recommend the practice generally.

Under five matched 100-epoch seeds without the contrastive objective (Table 3), real-plus-rendered reaches 4.46 ± 0.15 MAE against 4.44 ± 0.07 for real-only, so the gated stream does not confer a pooled MAE gain in that ablation; rendered-only remains worse at 4.66 ± 0.03. Under seven matched seeds with the full training objective, the pooled MAE of EchoStature



(a) Grad-CAM comparison.



(b) Counterfactual analysis.

Fig. 3: Interpretability analyses. (a) Grad-CAM attribution maps from the video-only reference model and EchoStature on subgroup-matched validation clips. (b) Counterfactual analysis of demographic conditioning. (i) Grad-CAM maps under controlled perturbation of sex, age bin, and BMI category while the echocardiographic video is held fixed. (ii) UMAP projections of the learned regressor representation colored by sex, age bin, and BMI category. Sex and age show partial separation, whereas BMI categories remain largely intermixed.

and the video-only reference are both 4.49, so a pooled architectural advantage is not established there either. The convex gate is therefore best read as a mechanism that can admit complementary rendered information without treating rendered video as a substitute for real echocardiography, rather than as a source of a confirmed pooled gain.

Demographic attributes contribute auxiliary context rather than dominant signal. Zeroing the video representation while retaining the demographic embedding raises MAE from about 4.49 to 10.69 on a representative checkpoint, and perturbing the demographic vector at inference changes MAE by at most about 0.04 across seven seeds. Predictions consequently remain stable under controlled demographic perturbation, indicating that the model does not substitute metadata for visual evidence of ventricular mechanics. Attribution maps nonetheless shift under the same perturbations, so demographic context influences which spatial evidence the model weights without appreciably altering the resulting estimate. This dissociation between attribution sensitivity and prediction stability merits closer study, since attribution

Table 4: Subgroup MAE for EchoStature on the validation set (mean ± SD over 7 seeds).

Subgroup	**MAE↓**	**Count**
Sex: female	4.23 ± 0.07	653
Sex: male	4.68 ± 0.07	905
Age: 0–1	5.79 ± 0.20	132
Age: 2–5	3.89 ± 0.14	220
Age: 6–10	3.94 ± 0.13	394
Age: 11–15	4.53 ± 0.06	480
Age: 16–18	4.97 ± 0.12	332
BMI: underweight	4.38 ± 0.11	749
BMI: normal	4.44 ± 0.06	543
BMI: overweight	5.17 ± 0.15	159
BMI: obese	4.51 ± 0.20	107

Table 5: Demographic conditioning ablation for EchoStature (mean ± SD over 7 seeds).

Mode	**MAE↓**	**MSE↓**	**RMSE↓**	**R^2 ↑**
Normal	4.49 ± 0.04	38.12 ± 0.91	6.17 ± 0.07	0.587 ± 0.010
Zero	4.53 ± 0.04	39.09 ± 1.04	6.25 ± 0.08	0.576 ± 0.011
Shuffle	4.49 ± 0.04	37.98 ± 0.84	6.16 ± 0.07	0.589 ± 0.009

alone would suggest a stronger demographic dependence than the predictions exhibit.

Several limitations constrain these conclusions. Evaluation uses a fixed clip-level partition that does not guarantee patient-disjoint generalization, and model selection uses the same validation set. Main-model and demographic-perturbation results are reported as mean ± SD over seven seeds, and the matched no-contrastive input ablation as mean ± SD over five seeds, all on a single cohort; generalization across acquisition protocols remains untested. The comparison contrasts demographic-aware gated fusion against video-only regression, and it therefore does not isolate the contribution of the rendered stream from that of the demographic embedding. A no-contrastive configuration was trained under the same 100-epoch budget and summarized over five seeds, but because that protocol still differs from the seven-seed full model in more than the contrastive weight alone, the contribution of that objective remains unisolated. BMI categories rely on fixed thresholds rather than pediatric percentiles, which limits the granularity of body-size conditioning, and a percentile-based or continuous body-surface-area encoding would likely provide a stronger signal.[11] Grad-CAM and counterfactual analyses probe behavior without establishing clinical reliability, and expert evaluation and uncertainty estimation remain necessary. The dark-cavity connected-component check on real–rendered pairs likewise remains an image-analysis measurement rather than clinical left-ventricle segmentation.

The most consequential direction is to reduce error where it is largest. The higher infant error may be related to fixed-length, fixed-resolution encoding applied to a population with high heart rates and small chambers, but those mechanisms were not directly tested. Adaptive temporal sampling or resolution scaled to body size is therefore a hypothesis for

later experiments, rather than an established remedy. Whether accuracy and demographic independence trade off in pediatric EF regression is a second open question, and answering it requires perturbation analyses applied across architectures rather than to a single model.

Future evaluation should use patient-disjoint training, validation, and test partitions, so that performance reflects generalization to patients not seen in training. Validation on larger external cohorts, preferably from more than one center and acquired under different protocols, is also needed to determine whether the age-group error pattern and the gated-fusion comparison hold beyond EchoNet-Pediatric.

5. Conclusion

We presented EchoStature, a pediatric EF regression model that fuses each real acquisition with a paired demographic-conditioned reconstruction through a learned gate. Under multi-seed evaluation the pooled difference from the video-only reference is negligible, and predictions remain stable under controlled demographic perturbation. Subgroup-stratified evaluation reveals larger variation across pediatric age groups than between architectures, with infants incurring the highest error. Pediatric EF regression therefore warrants subgroup-stratified reporting as standard practice, since pooled accuracy conceals precisely the population for which echocardiography is most often the only available assessment of ventricular function.

References

1. D. Ouyang, B. He, A. Ghorbani, N. Yuan and Others, Video-based ai for beat-to-beat assessment of cardiac function, *Nature* **580**, 252 (2020).
2. C. D. Reddy, L. Lopez, D. Ouyang, J. Y. Zou and B. He, Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients, *Journal of the American Society of Echocardiography* **36**, 482 (2023).
3. D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun and M. Paluri, A closer look at spatiotemporal convolutions for action recognition, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
4. M. Mokhtari, T. Tsang, P. Abolmaesumi and R. Liao, EchoGNN: Explainable ejection fraction estimation with graph neural networks, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, , Lecture Notes in Computer Science Vol. 13434 (Springer, 2022) pp. 360–369.
5. S. Thomas, A. Gilbert and G. Ben-Yosef, Light-weight spatio-temporal graphs for segmentation and ejection fraction prediction in cardiac ultrasound, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, , Lecture Notes in Computer Science Vol. 13434 (Springer, 2022)
6. H. Li, Y. Wang, M. Qu, P. Cao, C. Feng and J. Yang, EchoEFNet: Multi-task deep learning network for automatic calculation of left ventricular ejection fraction in 2D echocardiography, *Computers in Biology and Medicine* **156**, p. 106705 (2023).
7. H. Reynaud, A. Vlontzos, B. Hou, A. Beqiri, P. Leeson and B. Kainz, Ultrasound video transformers for cardiac ejection fraction estimation, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.
8. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone and E. Ferrante, Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, *Proceedings of the National Academy of Sciences* **117**, 12592 (2020).
9. E. Puyol-Antón, B. Ruijsink, S. K. Piechnik, S. Neubauer, S. E. Petersen, R. Razavi and A. P. King, Fairness in cardiac MR image analysis: An investigation of bias due to data imbalance in deep learning based segmentation, in *MICCAI*, , LNCS Vol. 129032021.
10. L. Seyyed-Kalantari, H. Zhang, M. B. McDermott, I. Y. Chen and M. Ghassemi, Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations, *Nature medicine* **27**, 2176 (2021).
11. L. Lopez, S. Colan, M. Stylianou, S. Granger, F. Trachtenberg, P. Frommelt, G. Pearson, J. Camarda, J. Cnota, M. Cohen *et al.*, Relationship of echocardiographic z scores adjusted for body surface area to age, sex, race, and ethnicity: the pediatric heart network normal echocardiogram database, *Circulation: Cardiovascular Imaging* **10**, p. e006979 (2017).
12. S. Lee, J. Kim, K. Kang and M. Kang, Explainable integrative bipartite graph convolutional neural network for predicting ejection fraction in echocardiography, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, , Lecture Notes in Computer Science Vol. 15971 (Springer, 2025) pp. 331–341.
13. T. Park, M.-Y. Liu, T.-C. Wang and J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
14. E. Perez, F. Strub, H. de Vries, V. Dumoulin and A. Courville, FiLM: Visual reasoning with a general conditioning layer, in *Proceedings of the AAAI Conference on Artificial Intelligence*, (1)2018.
15. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge and F. A. Wichmann, Shortcut learning in deep neural networks, *Nature Machine Intelligence* **2**, 665 (2020).

16. M. Mirza and S. Osindero, Conditional generative adversarial nets (2014).
17. D. Tran, L. Bourdev, R. Fergus, L. Torresani and M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.
18. H. Reynaud, A. Vlontzos, M. Dombrowski, C. Gilligan Lee, A. Beqiri, P. Leeson and B. Kainz, D'ARTAGNAN: Counterfactual video generation, in *MICCAI*, , LNCS Vol. 134382022.
19. R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. Williamson and F. Mahmood, Synthetic data in machine learning for medicine and healthcare, *Nature biomedical engineering* **5**, 493 (2021).
20. H. Reynaud, M. Qiao, M. Dombrowski, T. Day, R. Razavi, A. Gomez, P. Leeson and B. Kainz, Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis, in *International conference on medical image computing and computer-assisted intervention*, 2023.
21. T. Chen, S. Kornblith, M. Norouzi and G. Hinton, A simple framework for contrastive learning of visual representations, in *International conference on machine learning*, 2020.
22. P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu and D. Krishnan, Supervised contrastive learning, *Advances in neural information processing systems* **33**, 18661 (2020).
23. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
24. A. Ghorbani, D. Ouyang, A. Abid, B. He, J. H. Chen, R. A. Harrington, D. H. Liang, E. A. Ashley and J. Y. Zou, Deep learning interpretation of echocardiograms, *NPJ digital medicine* **3**, p. 10 (2020).
25. J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt and B. Kim, Sanity checks for saliency maps, *Advances in neural information processing systems* **31** (2018).
26. S. Wachter, B. Mittelstadt and C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the GDPR, *Harvard Journal of Law & Technology* **31**, 841 (2017).