This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

A Full Cross-Validation Results (Development)

Patient-level five-fold cross-validation is our development and model-selection protocol; the official validation split (Sec. 4.4) is the held-out report. Table 3 gives the complete per-configuration cross-validation underlying the label-noise-floor analysis (Sec. 4, Fig. 4): every architecture, loss, foundation encoder, and transfer variant ties the in-domain baseline within one fold standard deviation, and only the six-seed ensemble significantly improves on it.

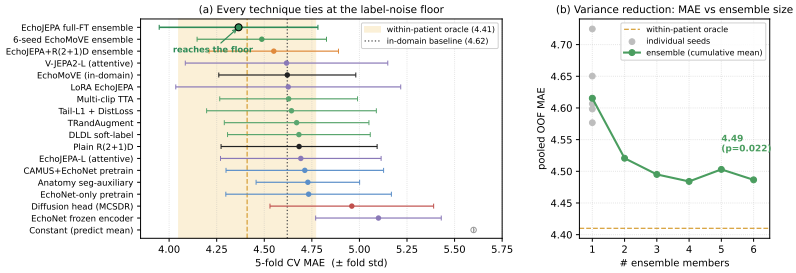


Fig. 4. (a) Cross-validated MAE (± fold std) for every configuration, ordered worst-to-best; the gold band is the oracle floor (4.41 ± 0.36) and the dotted line the in-domain baseline. The whole field—architectures, losses, foundation encoders, transfer, LoRA—collapses into a ≈0.2-MAE band; only the diffusion head and frozen-encoder transfer are reliably worse, and only the six-seed ensemble reliably better. (b) The six seeds scatter over 4.58–4.73 MAE but their cumulative average converges to 4.49: no single model escapes the floor, averaging does.

Table 3. Patient-level 5-fold CV on EchoRisk Task 1 (identical folds, seed 42). Std denotes fold-to-fold standard deviation. EchoMoVE and the plain R(2+1)D baseline are not statistically separable (paired t-test p=0.29, Wilcoxon p=0.44); the foundation-model, transfer, and single-model variants all tie the baseline within one standard deviation; the diffusion head is reliably worse, and the six-seed ensemble is the *only* configuration reliably better (paired p=0.022).

Method	CV MAE ± Std	Pearson r	RMSE
Constant training-mean (LVEF 64)	5.60 ± 0.00	—	7.18
R(2+1)D baseline (single-view, per-view avg) †	4.81	0.44	—
Plain R(2+1)D (exam-level: no MoE/MIL/contrastive)	4.68 ± 0.41	0.40	—
EchoMoVE (regularised MoE, no biomarker)	4.62 ± 0.36	0.43	—
+ Multi-clip test-time averaging	4.63 ± 0.36	0.42	—
+ DLDL soft-label head	4.68 ± 0.38	0.43	—
+ C-Mixup regression mixup	4.60 ± 0.37	0.43	—
+ Diffusion regression head (MCSDR)	4.96 ± 0.43	0.39	—
CAMUS+EchoNet pretrain, fine-tuned	4.71 ± 0.41	0.43	—
EchoNet-only pretrain, fine-tuned	4.73 ± 0.44	0.45	—
EchoNet pretrain, encoder frozen	5.10 ± 0.33	0.27	—
Anatomy seg-auxiliary (CAMUS ED/ES)	4.73 ± 0.27	0.41	—
EchoJEPA-L, frozen + attentive read-out	4.69 ± 0.42	0.40	—
V-JEPA2-L, frozen + attentive read-out	4.62 ± 0.53	0.40	—
EchoJEPA + R(2+1)D ensemble	4.55 ± 0.34	0.46	6.36
6-seed EchoMoVE ensemble ‡	4.49 ± 0.34	0.48	—

†Per-view baseline evaluated on the single 156-exam validation split. The exam-level *Plain R(2+1)D* row is its matched 5-fold CV counterpart—EchoMoVE with the MoE, masked MIL fusion, and contrastive regularisation all disabled—and isolates the contribution of those modules. ‡Mean of six independently seeded EchoMoVE models on identical folds—the *only* configuration that significantly improves on the single-model baseline (paired t-test p=0.022).

B Modular Ablation

Table 4. Single-split modular ablation (development val MAE, 156 exams) under the regularised training protocol. Differences ≤0.05 MAE are within noise; ≥0.30 MAE are reliable.

Configuration	Val MAE	r
EchoMoVE (all modules, no biomarker)	**4.451**	0.482
Biomarker fusion added	4.775	0.433
Mean fusion (MIL removed)	4.811	0.413
Contrastive loss removed	4.806	0.391
MoE replaced by single expert	4.704	0.472
Encoder stem frozen	4.854	0.425
Only 15 training epochs	4.876	0.379
EchoPrime encoder, 224-px, frozen	4.641	0.490
EchoPrime encoder, 224-px, fine-tuned	4.736	—

C Architecture and Training Hyperparameters

Table 5 lists every setting needed to reproduce the submitted EchoMoVE model; all values are fixed before cross-validation with no fold-specific tuning. Code and the inference Docker accompany the submission.

Table 5. Full configuration of the submitted six-seed EchoMoVE ensemble.

Component	Setting
Backbone	R(2+1)D-18, Kinetics-400 init, feature dim D=512
Input	2 views (A4C, A2C), 32 frames, 128→112 crop, grayscale→3ch
Frame sampling	one random 32-frame consecutive window (stride 1) in train; centre window at eval
Normalisation	Kinetics-400 mean/std
Clips per view (K)	1 (train and eval); K=3 evenly-spaced windows for the multi-clip TTA row
View-pool attention	masked additive attention: Lin(512,128)− tanh −Lin(128,1)
MoE head	3 experts, each Lin(512,128)−ReLU−Lin(128,1); linear gate
Expert bands	Gaussian centres $\{45, 60, 75\}$, σ=12
Projection	LayerNorm + Dropout 0.6
Loss	$\mathcal{L}_{\ell_1} + 0.1\,\mathcal{L}_{\text{con}} + 0.2\,\mathcal{L}_{\text{expert}}$
Optimiser	AdamW, lr 1×10^{-4}, weight decay 5×10^{-4}
Schedule	cosine, 2 warmup epochs, 25 epochs total, AMP, grad-clip 1.0
Batch size	16 exams/step (each exam ≤2 views)
Train augmentation	view-dropout 0.3, brightness/contrast jitter 0.2
Ensemble	6 seeds, predictions averaged at inference (init seed varied; folds and data fixed)
Checkpoint	best validation MAE per fold
Hardware/SW	1×NVIDIA RTX-PRO-6000 (sm_120), PyTorch 2.8/CUDA 12.8; ≈5.5 min/fold
Conformal	split CV+, score $\lvert y-\hat{y}\rvert$, $q_\alpha = \lceil (n+1)(1-\alpha)\rceil/n$ quantile, α=0.1

Preprocessing and augmentation. Each DICOM cine loop is decoded to grayscale, resized to 128^2, and a single 32-frame consecutive window (stride 1) is sampled—at a random start in training, the centre window at inference (short loops wrap around). The window is replicated to 3 channels and Kinetics-normalised. Train-time augmentation is a random 112^2 crop (centre crop at eval) plus multiplicative brightness and contrast jitter, each $\times[0.8, 1.2]$; we use *no* horizontal flip (A2C/A4C laterality is meaningful), rotation, or added noise. One of the two views is randomly masked out with probability 0.3 to train the MIL fusion to degrade gracefully.

MoE training and design choices. All K=3 experts receive a *soft-weighted* gradient every step—expert k's loss is scaled by its Gaussian membership $r_k(y)$ (Eq. 2), so there is no hard routing and no expert is starved; the linear gate is trained only through the final ℓ_1 on the gated prediction (no entropy regulariser or gate supervision, and we observed no gate collapse). The three band centres $\{45, 60, 75\}$ anchor the low/normal/high EF ranges spanning the label distribution; collapsing to a single expert costs +0.25 MAE on the development split (Table 4). Remaining choices are set by the single-split ablation (Table 4): dropout 0.6 (> 0.5, to curb overfitting on 630 exams), view-dropout 0.3, and the 128-d additive-attention bottleneck (a small, standard read-out width); λ_c=0.1, λ_e=0.2.

Algorithm 1 EchoMoVE training step (one exam batch)

Require: clips $\{x_v\}$, presence mask $\{m_v\}$, label y; shared encoder f_θ

1: **for** each present view v **do** $\mathbf{e}_v \leftarrow f_\theta(x_v)$ ▷ D=512 spatiotemporal embedding
2: **end for**
3: $s_v \leftarrow \mathbf{w}^\top \tanh(\mathbf{W}\mathbf{e}_v)$; $\alpha_v \leftarrow m_v e^{s_v} / \sum_i m_i e^{s_i}$; $\mathbf{e} \leftarrow \sum_v \alpha_v \mathbf{e}_v$ ▷ masked MIL fusion
4: $g \leftarrow \mathrm{softmax}(\mathbf{W}_g \mathbf{e})$; $\hat{y} \leftarrow \sum_k g_k \, \phi_k(\mathbf{e})$ ▷ gated MoE prediction
5: $\mathcal{L} \leftarrow \mathcal{L}_{\ell_1}(\hat{y}, y) + \lambda_c \mathcal{L}_{\mathrm{con}}(\mathbf{e}, y) + \lambda_e \sum_k r_k(y) \, \mathcal{L}_{\ell_1}(\phi_k(\mathbf{e}), y)$
6: back-propagate $\mathcal{L}$; update with AdamW

Inference. A test exam's present views are encoded, fused (Eq. 1), and passed through the gated MoE to a single LVEF $\hat{y}$; the six independently-seeded models' $\hat{y}$ are then averaged. Notation is summarised in Table 6.

Table 6. Notation.

Symbol	Meaning
x_v, m_v	clip and binary presence mask for view $v \in \{\text{A4C, A2C}\}$
$\mathbf{e}_v$, $\mathbf{e}$	per-view embedding; fused exam embedding ($\in \mathbb{R}^{512}$)
α_v	masked view-attention weight (Eq. 1)
ϕ_k, g_k	expert-k regressor; gate weight for expert k
$r_k(y)$	Gaussian band membership of label y for expert k
$\mathcal{L}_{\ell_1}, \mathcal{L}_{\text{con}}, \mathcal{L}_{\text{expert}}$	ℓ_1 MAE; label-distance regulariser; expert-specialisation term
λ_c, λ_e	loss weights (0.1, 0.2)

D EchoMoVE vs. EchoJEPA: Resource Comparison

Table 7 compares the deployable EchoMoVE against the best-accuracy fully-fine-tuned EchoJEPA on compute footprint. All figures are measured on this work's hardware (1×NVIDIA RTX-PRO-6000, PyTorch 2.8/CUDA 12.8): training time from the per-fold training logs; per-process peak VRAM via `torch.cuda.max_memory_allocated`; inference latency benchmarked GPU-only (batch 1, cached clips, excluding DICOM decode). EchoMoVE is an order of magnitude smaller and ≈5.5× faster at inference for a ≈0.1 MAE difference.

Table 7. Per-model resource comparison (single model; the deployed system averages six seeds, multiplying inference cost ×6). Training VRAM differs mainly because the models train at different batch sizes (16 vs 4).

	EchoMoVE (R(2+1)D)	EchoJEPA full-FT (ViT-L)
Parameters	31.6 M	308.7 M
Checkpoint size (`best.pt`)	121 MB	1178 MB
Training time	5.5 min/fold (13.2 s/ep ×25)	10.7 min/fold (21.5 s/ep ×30)
Training peak VRAM	15.6 GB (batch 16)	7.2 GB (batch 4)
Inference latency	8.4 ms/exam	46.4 ms/exam
Inference peak VRAM	0.59 GB	2.50 GB
Val MAE (6-seed ensemble)	4.67	**4.56**

E External Transfer and Foundation-Encoder Study

We evaluated whether supervised LVEF pretraining on larger external corpora could overcome the data-scarcity constraint. This is categorically distinct from using frozen foundation-model features: we pretrain our own R(2+1)D backbone on the external LVEF task and subsequently fine-tune the complete EchoMoVE pipeline on EchoRisk.

EchoNet-Dynamic. We pretrain on the full EchoNet-Dynamic dataset (10,030 A4C cine loops; 12.6% with EF < 40) using the identical preprocessing pipeline

as EchoRisk (128→112 crop, Kinetics normalisation, 32 frames). The pretrained Stage-A encoder reproduces the EchoNet paper on its own held-out test set (MAE 4.13, R^2 0.79 vs. reported 4.05–4.22 / 0.79–0.81), confirming the validity of the pretraining. After fine-tuning EchoMoVE (encoder learning-rate multiplier 0.3), the CV MAE is 4.73 ± 0.44—a tie with the in-domain baseline. Freezing the encoder degrades performance to 5.10 ($r = 0.27$). The monotonic relationship—the more of the external prior that is preserved, the worse the result—indicates that A4C-only, single-centre EchoNet features are misaligned with the multi-centre, two-view EchoRisk distribution.

CAMUS. To address the A4C-only limitation, we include the CAMUS dataset [10] (500 patients, A2C and A4C, 46% with EF < 45, GE-dominant vendor). The combined CAMUS+EchoNet Stage-A achieves EchoNet-validation MAE 4.39. After fine-tuning, the CV MAE is 4.71 ± 0.41—still a tie in global MAE. However, Table 8 shows a clear improvement in low-LVEF detection, motivating the re-analysis in Sec. 4.7.

Frozen foundation encoders. EchoPrime at its native 224-px normalisation achieves val MAE 4.64 (frozen) and 4.74 (fine-tuned), both worse than the in-domain model despite adequate preprocessing. CoReEcho, which achieves 3.90 on EchoNet, degrades to 4.95 frozen and 5.09 fine-tuned on EchoRisk. The consistent pattern across both architectures confirms a domain-shift effect: models trained on single-view, single-centre corpora with precise Simpson's-method labels do not transfer zero-shot to a multi-centre, two-view, visually assessed cohort. **JEPA encoders and the read-out.** Self-supervised JEPA video encoders—generic V-JEPA2 [3] and the echo-specific EchoJEPA [13]—emit patch tokens rather than a pooled embedding, and we find the *read-out* is decisive: mean-pooling the tokens collapses the EF signal (EchoRisk r=0.10; EchoNet r=0.51), whereas a small trainable attentive pooler recovers it (EchoRisk r=0.40; EchoNet r=0.86). With the correct attentive read-out, frozen EchoJEPA-L attains CV MAE 4.69 ± 0.42—the best of any single foundation encoder, yet still a statistical tie with the in-domain baseline (paired p=0.53); with attentive read-out even generic V-JEPA2-L ties the baseline (4.62 ± 0.53, up from 4.81 under mean-pooling). We validate the EchoJEPA checkpoint on its home benchmark: a frozen attentive probe reproduces the published EchoJEPA-L on EchoNet-Dynamic (our TEST MAE 4.66, r=0.86 vs. their reported 4.85), confirming the encoder is genuinely strong—the EchoRisk tie is a property of the target cohort, not a weak model. (Their headline 3.97 is the ViT-Giant, whose weights are not publicly released.) Finally, ensembling EchoJEPA with the in-domain R(2+1)D gives the lowest CV MAE we observe (4.55 ± 0.34) and the highest correlation (r=0.46), improving on the baseline in four of five folds—but the gain remains non-significant (p=0.34). Even the best constructible configuration sits at the label-noise floor.

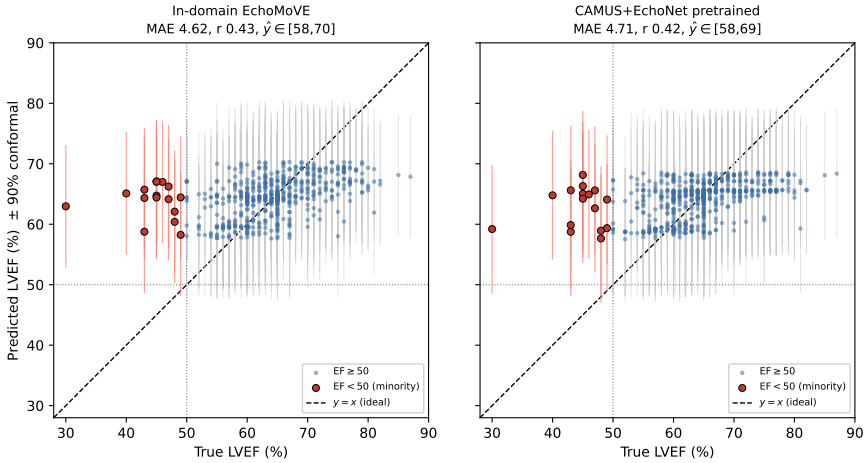


Fig. 5. True vs. predicted LVEF on pooled out-of-fold predictions, with 90% split-conformal intervals (vertical bars); the low-LVEF minority (EF<50) is highlighted in red. Both models compress their predictions into a narrow 57–71 band: low-LVEF exams (true 30–50) are predicted near 60, and even the 90% conformal interval fails to reach the true value—visually proving the 0% interval coverage below LVEF 45. Tail-rich CAMUS pretraining (right) does *not* decompress the predictions; its benefit is confined to ranking (detection AUC, Table 8), reinforcing that the limitation is the absolute-MAE axis itself, not a weak model.

F Low-LVEF Detection and Conformal Analysis

Global MAE is an inadequate summary for this task: on a distribution massed at LVEF ≈64, predicting the mean is ℓ_1-optimal yet blind to the low-LVEF minority that defines cardiotoxicity. We therefore re-score on low-LVEF detection AUC over pooled out-of-fold predictions, treating the regression output as a ranking.

Table 8. Low-LVEF detection AUC with 95% confidence intervals from a patient-clustered bootstrap (2000 resamples). n positive denotes the number of positive exams in the pooled out-of-fold set.

Model	EF < 50 AUC (95% CI)	EF < 45 AUC (95% CI)	r
EchoMoVE, in-domain	0.56 (0.43–0.68)	0.61 (0.40–0.95)	0.43
EchoJEPA, frozen + attentive	0.61 (0.48–0.75)	0.45	0.40
CAMUS+EchoNet pretrain	0.63 (0.49–0.74)	**0.71** (0.54–0.95)	0.42
EchoJEPA + task-query decoder	**0.67** (0.52–0.78)	—	0.39
n positives	16	5	—

Two results follow (Table 8). First, CAMUS+EchoNet pretraining raises detection AUC by 0.07–0.09 while global MAE stays flat (4.71 vs. 4.62): the tail-rich CAMUS distribution (46% below EF 45) supplies low-LVEF visual features the EchoRisk training set cannot. Second, repurposing the frozen EchoJEPA encoder with a small query-decoder (learnable view/task tokens cross-attending the patch map, trained with a hybrid ℓ_1/low-EF/ranking objective) does not improve global MAE (4.73) but *raises* EF <50 detection to AUC 0.67, the best in our study. Both make the recurring pattern explicit: foundation-model capacity on this cohort converts not into lower absolute error—which label noise bounds—but into better ranking of the minority the task targets. (With only 5–16 positive exams the CIs span to chance, so these are consistent trends, not conclusive results—a minority-class sample-size limit, not a modelling failure.)

Conformal intervals expose the same mismatch structurally: overall coverage is 0.897, but per stratum it is 0.95 for LVEF >55, 0.27 for 45–55, and **0.00** below 45, because the model's predictions compress to a 57–71 band and never go below 55 (Appendix E, Fig. 5). This argues for selective-prediction deployment over raw regression: return an automated estimate only when the *lower* 90% conformal bound exceeds 50, and otherwise flag the exam for a manual Simpson's read. The rule never issues a confident normal reading for a true low-LVEF exam, converting miscalibration into safe abstention exactly where the model is unreliable.

G Test-Set Generalization Methods

On the best base (fully fine-tuned EchoJEPA) we additionally evaluated the weight-space and robustness methods most associated with distribution-shift generalization, to test whether the floor can be pushed further. Under matched 5-fold CV, *none* lowered MAE below the 6-seed ensemble (4.37): SWA (dense trajectory weight averaging) 4.52; surgical fine-tuning (last-4 blocks only) 4.60; sharpness-aware minimisation (SAM) 4.97 and tail-weighted BSAM 4.82; test-time augmentation neutral (4.57 val); and uniform *model soups* of the six fully-fine-tuned seeds *failed* (5.20)—fully fine-tuned ViT-L seeds drift to distinct loss basins, so their weights do not average (predictions still must be ensembled). The lone exception is on the secondary axis: SAM and BSAM raise low-LVEF detection AUC to 0.666/0.656 (vs 0.635), trading ≈0.3 MAE for it. This exhausts the standard generalization toolkit and reinforces the label-noise-floor finding: variance reduction over a strong base is the only lever that improves MAE. All MAE/RMSE/r figures in this paper are computed with the official `echorisk-score` evaluator.