General vs. Echo-Specific Video Foundation Models for Echocardiographic Risk

Rene Bidart[1] and Shrimanti Ghosh[2]

[1] Independent Researcher
renebidart@gmail.com
[2] Department of Radiology and Diagnostic Imaging, University of Alberta, Canada
shrimant@ualberta.ca

Abstract. The EchoRisk 2026 challenge defines three tasks on apical echocardiography video from a cardio-oncology cohort: left-ventricular (LV) ejection fraction (EF) regression (Task 1), dysfunction classification defined by global longitudinal strain (Task 2), and prediction of early cardiotoxicity from a single pre-treatment scan (Task 3), each scored per examination. We use self-supervised video encoders as frozen feature extractors and train an attentive-pooling probe on top. Under one protocol, we compare general-purpose and echo-pretrained encoders. On Task 1, adding public EchoNet-Dynamic EF labels to the probe's training data reduces the mean absolute error (MAE) of a general video encoder from 4.91 to 4.56, matching an echo-pretrained encoder and improving on the organisers' supervised baseline of 5.02. Fine-tuning the encoder does not improve on the frozen probe at this label scale. Content hashing also revealed repeated cines across released training and validation keys, sometimes with conflicting labels; we remove these records before all reported experiments. Quantitative results are reported for Task 1 only, on the validation split that was also used for model selection, and may therefore be optimistic; we make no hidden-test claim.

Team: rbid.

Keywords: Echocardiography · Ejection fraction · Cardiotoxicity · Video foundation models · Self-supervised learning · Transfer learning.

1 Introduction

Echocardiography is the standard tool for monitoring cardiac function in patients receiving cancer therapy. Left-ventricular ejection fraction (EF) is the usual functional measure, but it falls late in the course of cardiac injury. Cardio-oncology guidelines recommend global longitudinal strain (GLS) for surveillance because strain declines before EF, and a relative drop in GLS can indicate subclinical injury while EF is still preserved [1, 2]. A model that predicts only EF will miss part of the population that early surveillance is meant to catch.

The EchoRisk 2026 challenge is organised around this distinction [3]. It defines three tasks on apical four-chamber (A4C) and two-chamber (A2C) cine

video from a multi-site cardio-oncology cohort: EF regression (Task 1), GLS-defined LV dysfunction (Task 2), and prediction of early cardiotoxicity from a single pre-treatment scan (Task 3). The tasks are scored independently and per examination, against a hidden test set with no development leaderboard.

We approach the three tasks by transfer learning from video foundation models. Our submission uses a base V-JEPA-2 ViT-L encoder [4], a general-purpose self-supervised video model, as a frozen feature extractor with an attentive-pooling probe. We compare it with echo-pretrained encoders [5], test whether fine-tuning the encoder helps, and test whether additional public EF labels in the probe's training data help.

Contributions. (1) We benchmark general and echo-pretrained video encoders on all three EchoRisk tasks under one frozen-probe protocol, reported on the official per-exam metric. (2) On Task 1, adding public EchoNet-Dynamic EF labels to the probe's training data reduces a general encoder's per-exam MAE from 4.91 to 4.56, matching an echo-pretrained encoder at 4.55. (3) Fine-tuning the encoder does not improve on the frozen probe with this amount of labelled data. (4) We identify and remove duplicated cines in the released data that otherwise cause leakage between the train and validation splits.

We report validation-set results only, and quantitative results for Task 1 only: the Task 2 and 3 results from the reviewed version of this paper are withdrawn, with the reasons stated, in Sect. 4. The final offline inference system may continue to evolve after the paper deadline under the challenge rules and is evaluated separately on the hidden set.

2 Data and Preprocessing

Cohort and splits. EchoRisk provides 2,159 A4C and A2C cine videos from 422 patients, with up to five timepoints per patient. The organisers give patient-disjoint train, validation, and test splits (about 56/14/30%); the test split is hidden and there is no development leaderboard, so all development here uses the validation split. We construct the development manifests using the per-patient challenge identifier. A patient contributes several timepoints, so a clip-level split would place one patient in both train and validation. We do not use the DICOM internal `PatientID` because its prefix encodes the acquisition site rather than identifying the subject. The released identifier keeps train and validation subject-disjoint (verified zero overlap).

Duplicated imaging and deduplication. Content-hashing the released cines shows that the same physical video is stored under several (patient, timepoint) keys, sometimes under different patient identifiers and across the train and validation splits, and often with conflicting labels (for example, pixel-identical clips labelled EF 82 in train and EF 59 in validation). The source DICOM pixel arrays were independently verified as identical, so this is a defect in the release rather than a transcoding artefact. A pipeline that skips deduplication therefore leaks training content into validation. Before every reported experiment, we compute a content

Table 1. Corrected development manifests after content-hash deduplication. Clip and patient counts describe processed data; the last column is the number of officially scored exams, which the scorer requires in full even where processing leaves an exam without usable video (Sect. 3). For Task 3, the positive-exam count is in parentheses.

Task	Processed clips	Train / val patients	Official exams
1 (EF regression)	1,440	236 / 59	156
2 (LV dysfunction)	1,465	237 / 59	156
3 (early cardiotoxicity)	386	156 / 37	39 (20 positive)

hash of each encoded video clip. Groups with conflicting labels are dropped in full. When identical clips share one label, we keep a single copy and prefer the training copy if the group spans both splits. A regression test enforces that no kept content hash crosses the split. This removes 63 of the 1,503 Task 1 clips, 38 of 1,503 for Task 2, and 6 of 392 for Task 3; the processed Task 1 validation panel then covers 152 of the 156 officially scored exams. Deduplication has been part of the data pipeline since its first version, so it precedes every result reported in this paper, and no pre-deduplication task metric was ever computed. Table 1 gives the corrected development manifests. To keep the submitted and competition-stage Task 1 runs on the same scale, we retain the frozen pre-deduplication training statistics $\mu = 64.2155$ and $\sigma = 7.1535$. They were estimated from training per-view records only, without validation labels.

Missing calibration. The released DICOMs do not carry the temporal or spatial calibration tags (`FrameTime`, `CineRate`, `PixelSpacing`); only `NumberOfFrames` is present. No true frame rate or pixel scale is available at inference, which rules out methods that assume a calibrated frame rate or millimetre-per-pixel scale.

Preprocessing. Each cine is decoded to grayscale, cropped to its ultrasound sector (estimated from the temporal max-projection of the clip, which exposes the static fan boundary), resized to 224×224, and encoded at a fixed nominal 16 fps. At training and inference we sample K clips per video ($K = 2$ by default) and pool them. The same preprocessing is used for development and for the eventual offline container.

External data. We use only public datasets and published model weights, listed in Sect. 5. EchoNet-Dynamic [6] is used to select the EF backbone and as additional EF training data, with the encoder kept frozen. We also run the public weights of an echo latent-predictive model (EchoJEPA) [5], a multi-task echo model (PanEcho) [8], and a vision-language echo model (EchoPrime) [9] under our own frozen-probe protocol.

3 Methods

Backbone. Our default encoder is a base Meta V-JEPA-2 ViT-L, a general-purpose video model that is not echo-pretrained, loaded through an open echo-

adapted training fork [5]. We compare it with a downloadable echo-pretrained latent-predictive encoder (EchoJEPA-L and its 2.1 variant) and, as external task models, PanEcho and EchoPrime.

Frozen attentive probe. Our submitted method keeps the encoder frozen and trains an attentive-pooling probe on the frozen features, following the multi-head probing protocol used for V-JEPA-style video models [4, 5]. Task 1 uses a regression head; Tasks 2 and 3 use binary classification heads.

Fine-tuning. We test whether unfreezing the encoder helps. Naive fine-tuning updates the whole encoder end-to-end. We also run linear-probe-then-fine-tune with layer-wise learning-rate decay (LP-FT + LLRD) [10]. LP-FT was proposed to reduce feature distortion: at low label counts, an under-converged head can move useful pretrained features off-distribution during fine-tuning.

Additional EF training data (Task 1). Task 1 has a few hundred labelled patients. We add public EchoNet-Dynamic EF clips to the probe's training set, with the encoder frozen, so that only the probe sees the extra data. The EchoNet labels are z-scored using the frozen EchoRisk training values $\mu = 64.2155$ and $\sigma = 7.1535$ so both sources share a scale. These values were computed from pre-deduplication training per-view records; validation labels were not used. This tests whether extra public EF labels can substitute for echo-specific pretraining.

Per-exam scoring. The official metric aggregates the A4C and A2C views for each (patient, timepoint) examination. We average the per-clip predictions to the exam level (mean EF for Task 1, mean positive-class probability for Tasks 2 and 3) and report per-exam numbers throughout. Model selection and reporting use per-exam predictions. Where an examination has no usable processed video, its official scoring row receives a fixed fallback value (the training-mean EF of 64.2155 for Task 1 and 0.5 for Tasks 2 and 3).

4 Experiments and Results

Protocol. Development uses the validation split: Task 1 is selected on validation MAE, and Tasks 2 and 3 on validation area under the receiver operating characteristic curve (AUROC), using the same set for selection and reporting (an optimistic but documented choice given the hidden test). For Task 1, the frozen-probe and fine-tuning results and the results using additional EF data are three-seed means; the remaining rows are single-seed. As a reference we list the organisers' published supervised baseline, an R(2+1)D + LSTM network [3].

Encoder. On EchoNet-Dynamic, our frozen base V-JEPA-2 probe reaches EF MAE 4.10. The generic V-JEPA 2.1 variant reaches 5.19, so we use base V-JEPA-2 as the general-video default. On the challenge data (Table 2) the downloadable echo-pretrained encoder leads Task 1, as expected from domain-matched pre-training on the small challenge cohort. The published EchoNet EF numbers for

Table 2. Task 1 EF regression on the processed, deduplicated EchoRisk validation panel ($n = 152$ exams from 59 patients; official per-exam metric; lower MAE is better; $K = 2$). These development rows do not use inference fallback for the four additional official exams. Rows marked "3-seed" are three-seed means; the others are single-seed. "+ EchoNet EF" adds public EchoNet-Dynamic EF clips while the encoder remains frozen. The organisers' baseline is their separately published validation result and was not recomputed on this panel.

Method	per-exam MAE ↓
Training-mean predictor (trivial baseline)	5.38
Organisers' baseline: R(2+1)D + LSTM (val)	5.02
Base V-JEPA-2, frozen probe (3-seed)	4.91
Base V-JEPA-2, full fine-tune (3-seed)	4.77
Base V-JEPA-2, LP-FT + LLRD (3-seed)	4.91
Base V-JEPA-2, frozen probe + EchoNet EF (3-seed)	4.56
Base V-JEPA-2, A4C/A2C pooled frozen probe	**4.52**
EchoJEPA-L (echo-pretrained), frozen probe (3-seed)	4.63
EchoJEPA-L, frozen probe + EchoNet EF (3-seed)	4.55
EchoPrime, frozen probe	5.30
PanEcho, frozen probe	5.37

EchoPrime, PanEcho, and EchoJEPA come from different splits and protocols and are not a comparison we ran; our direct comparison is on the challenge tasks below.

Task 1: EF regression. Table 2 reports per-exam MAE. Fine-tuning does not improve on the frozen probe. Naive fine-tuning lowers the mean slightly (4.77 vs 4.91), but the difference does not survive resampling or seed variation: paired exam-level bootstrap intervals include zero in every seed, and the sign changes across seeds. LP-FT + LLRD matches the frozen probe (4.91). Adding public EchoNet EF labels lowers the general encoder's MAE from 4.91 to 4.56, with each seed's bootstrap interval for the improvement excluding zero, and lowers the echo-pretrained encoder's MAE only slightly (4.63 to 4.55). The two backbones are level once the extra EF labels are added. The best single point estimate is an A4C/A2C pooled frozen probe at 4.52. This row keeps the encoder frozen and trains the same attentive probe on a single pooled training set that adds public EchoNet-Dynamic (A4C) and CAMUS (A2C and A4C) [7] EF-labelled cines to the EchoRisk training clips, with every label z-scored by the frozen training statistics of Sect. 3. Model selection stays on the EchoRisk validation split, and per-exam scoring is unchanged. All V-JEPA-2 and EchoJEPA-L configurations achieve a lower validation MAE than the organisers' supervised baseline of 5.02; the EchoPrime and PanEcho probes (5.30 and 5.37) improve only on the training-mean baseline of 5.38. The validation EF values occupy a narrow range, so differences between the principal frozen-probe configurations are small in absolute terms.

Tasks 2 and 3: withdrawn reviewed rows. The classification rows from the submitted methodology are not reused as evidence: their checkpoints were selected using accuracy rather than the challenge AUROC, and an earlier Task 3 pipeline resolved some records to cine videos belonging to the wrong task, which also restricted the reviewed Task 3 evaluation to 32 of the 39 official exams. The corrected competition-stage systems cover all three tasks and remain within the same video-model and per-exam aggregation framework: fixed per-exam ensembles that combine the model families above with task-specific temporal ensembles, selected on public validation. They are evaluated separately through the organisers' offline container process, and this paper reports no results for them.

5 Discussion and Conclusion

Our submission benchmarks frozen video foundation models on EchoRisk. On Task 1, adding public EF labels to a general video probe matches a downloadable echo-pretrained encoder (4.56 vs 4.55 MAE) and achieves a lower validation MAE than the organisers' supervised baseline of 5.02 (Table 2); fine-tuning the encoder does not help. At a few hundred labelled patients, additional public labels are more useful than adapting the encoder in this experiment. The competition-stage systems extend the submitted method with fixed, task-specific per-exam ensembles that supersede the withdrawn reviewed results. GLS-defined dysfunction and early cardiotoxicity remain plausibly harder for frozen global representations: GLS tracks finer myocardial motion than EF, and Task 3 offers far less development data. With the reviewed classification rows withdrawn, however, this paper provides no controlled encoder comparison on those endpoints and cannot isolate these effects.

Limitations. All numbers are on the validation set, which also drives model selection, so they may overstate hidden-test performance, especially on Task 1. The key contrasts are supported by paired exam-level bootstrap checks, but the table reports point estimates; future versions of this work should include confidence intervals for the important MAE differences. The released train and validation split is close to single-vendor (about 99% GE Vivid systems on the data we developed on), while the challenge cohort is multi-vendor, so our development set under-represents the acquisition variety of the hidden test. We make no test-set claim in this paper.

External resources. We use pretrained weights for Meta V-JEPA-2 ViT-L and the generic V-JEPA 2.1 variant; public EchoJEPA weights (V-JEPA-2 continued on a large echo corpus, ViT-L); PanEcho; EchoPrime; and Kinetics-400 initialization for the R(2+1)D temporal models. We use EchoNet-Dynamic, under the Stanford research-use agreement, for backbone selection and additional EF training data, and the public CAMUS dataset [7] as additional EF training data for the A4C/A2C pooled row. These resources are publicly accessible and remain subject to their respective licences and access terms.

Ethics and consent. This work is a secondary analysis of the de-identified EchoRisk challenge release. The challenge technical report states that each participating site obtained ethical approval under its national and local requirements and that patients who withdrew consent before data transfer were excluded. We did not recruit participants or access identifying data [3].

Conclusion. Frozen video foundation models with a small probe are a practical Task 1 baseline. In this experiment, public EF labels remove most of the gap between general and echo-pretrained encoders, while full encoder fine-tuning does not improve on the frozen probe. Fixed per-exam ensembles carry these model families into the three competition-stage systems.

Acknowledgements

This work was supported by the European Union Horizon 2020 CARDIOCARE project under grant agreement No. 945175. LLM-assisted tools were used for code and manuscript editing; the authors verified the experimental design, citations, scientific claims, and final text.

Disclosure of Interests

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

References

1. Lyon, A.R., et al.: 2022 ESC Guidelines on cardio-oncology developed in collaboration with the EHA, ESTRO and IC-OS. European Heart Journal **43**(41), 4229–4361 (2022)
2. Thavendiranathan, P., et al.: Use of myocardial strain imaging by echocardiography for the early detection of cardiotoxicity in patients during and after cancer chemotherapy: a systematic review. Journal of the American College of Cardiology **63**(25), 2751–2768 (2014)
3. Kalliatakis, G., et al.: EchoRisk: a multicentre echocardiography dataset and benchmark for cardio-oncology. arXiv:2607.01039 (2026)
4. Assran, M., et al.: V-JEPA 2: self-supervised video models enable understanding, prediction and planning. arXiv:2506.09985 (2025)
5. Munim, A., et al.: EchoJEPA: a latent predictive foundation model for echocardiography. arXiv:2602.02603 (2026)
6. Ouyang, D., et al.: Video-based AI for beat-to-beat assessment of cardiac function. Nature **580**(7802), 252–256 (2020)
7. Leclerc, S., et al.: Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. IEEE Transactions on Medical Imaging **38**(9), 2198–2210 (2019)
8. Holste, G., et al.: Complete AI-enabled echocardiography interpretation with multitask deep learning. JAMA **334**(4), 306–318 (2025). doi:10.1001/jama.2025.8731

9. Vukadinovic, M., et al.: Comprehensive echocardiogram evaluation with view primed vision-language AI. Nature **650**, 970–977 (2026). doi:10.1038/s41586-025-09850-x
10. Kumar, A., et al.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (ICLR) (2022)