MELBA Journal Machine Learning for Biomedical Imaging

Learning from Scarce Labels: Multi-View Echocardiography for Ejection Fraction Prediction

Zhiyuan **Gao** [1], Dominic **Yurk** [2], Yaser S. **Abu-Mostafa** [1]

1 Electrical Engineering Department, California Institute of Technology, Pasadena, CA 91125, USA
2 Asari AI, San Francisco, CA 94131, USA

Abstract

We present, to the best of our knowledge, the first publicly available resource for predicting left ventricular ejection fraction (EF) from parasternal long-axis (PLAX) echocardiography. Because no PLAX–EF datasets previously existed, our work focuses on an innovative data generation strategy to overcome this scarcity. By leveraging a time-based correlation between clinical notes and echocardiographic videos, combined with fine-tuning view classifiers and proxy labeling, we created a labeled dataset of over 25,000 PLAX videos. This enables us to train the first reproducible PLAX EF model, achieving a mean absolute error (MAE) of 6.86%. Given that apical four-chamber (A4C) methods, the clinical standard, report MAE values of 6%-7%, our results demonstrate that EF estimation from PLAX views is both feasible and clinically relevant. This surpasses the performance of existing methods and provides a clinically relevant solution for situations where apical views may not be feasible. Going further, we demonstrate that combining PLAX and A4C predictions via simple unweighted late fusion improves both single-view baselines to a 6.37% MAE, underscoring the value of multi-view integration. To promote continued research, we release the dataset labels, trained models, and runnable demos on GitHub, Hugging Face, and Google Colab: `https://github.com/Jeffrey4899/PLAX_EF_Labels_202509`.

Keywords

Apical four-chamber (A4C), Echocardiography, Ejection fraction, Multi-view fusion, Parasternal Long-Axis (PLAX), Proxy labeling, Scarce data, Video view classification.

Article informations

`https://doi.org/10.59275/j.melba.2026-8194`

Volume 2026, Received: 2025-12-30, Published yyyy-m2-d2

Corresponding author: zgao2@caltech.edu

Special issue: Medical Imaging with Deep Learning (MIDL) 2025

Guest editors: Lisa Koch, Ronald M. Summers, Chen Chen, Yan Zhuang

1. Introduction

Cardiovascular disease remains a leading cause of mortality worldwide (Tsao et al., 2023). Transthoracic echocardiography is a ubiquitous, non-invasive imaging modality for cardiac assessment, and left ventricular ejection fraction (EF) is one of the most clinically actionable summary measures of systolic function. EF informs diagnosis, treatment selection, and longitudinal monitoring across conditions such as heart failure and cardiomyopathies.

In routine clinical reporting, EF is most commonly derived from apical views (notably apical four-chamber, A4C), which provide favorable geometry for standardized volumetric measurements (Section 2). However, high-quality apical acquisitions can be difficult in real-world settings, including poor acoustic windows, limited patient positioning, and time-constrained bedside/point-of-care scans. In contrast, the parasternal long-axis (PLAX) view is routinely acquired and often easier to obtain in practice, and clinicians frequently use PLAX for rapid qualitative assessment of LV function (Russell et al., 2022). Despite its ubiquity, there is no widely adopted protocol for reporting EF directly from PLAX cine videos, and consequently there is limited *public* supervision linking PLAX to EF at scale.

This creates a major bottleneck for machine learning: while large public datasets exist for A4C EF regression (e.g., EchoNet-Dynamic (Ouyang et al., 2020)), public PLAX datasets are typically labeled for *other* targets (e.g., LVH-related phenotypes in EchoNet-LVH (Duffy et al., 2022)), and large clinical archives (e.g., MIMIC-IV-Echo) provide videos without view labels (Gow et al., 2023; Goldberger et al., 2000). Prior PLAX EF efforts are therefore either

proprietary/closed or rely on indirect measurement pipelines; reproducible PLAX EF benchmarks remain scarce (reviewed in Section 3).

Overview. We address the PLAX label scarcity problem by constructing, to the best of our knowledge, the first publicly released resource that associates PLAX cine videos with EF labels at scale using only publicly accessible datasets. Our pipeline combines: (i) robust video view classification to mine A4C and PLAX from large unlabeled repositories, (ii) time-based correlation between studies and clinical notes to form an independent, note-derived ground-truth cohort, and (iii) *proxy supervision* in which A4C-based EF predictions provide study-level labels for PLAX training when direct PLAX EF annotations are unavailable. Beyond single-view modeling, we demonstrate that simple study-level late fusion of A4C and PLAX predictions yields consistent gains, supporting multi-view EF estimation.

Contributions. Concretely, we make the following contributions:

- **A public PLAX–EF resource.** We release label files and end-to-end artifacts that enable reproducible PLAX EF research using MIMIC-IV resources, including instructions for locating the corresponding samples under the data-use agreement (Gow et al., 2023; Johnson et al., 2023; Goldberger et al., 2000).
- **A reproducible PLAX EF benchmark.** Training on large-scale proxy-supervised PLAX data and evaluating on a note-derived ground-truth cohort, our PLAX ensemble achieves **6.86% MAE** at the study level.
- **Multi-view EF estimation.** Unweighted late fusion of A4C and PLAX predictions improves both single-view baselines, achieving **6.37% MAE** and Pearson correlation 0.709 on studies containing both views.
- **Open artifacts for adoption and extension.** We release runnable demos and pretrained models (GitHub, Hugging Face, and Colab) to support independent validation and downstream research (Gao et al., 2025b).

Extensions over the MIDL 2025 proceedings version. As required for MELBA submissions extending a conference paper, this manuscript substantially expands the MIDL 2025 proceedings version (Gao et al., 2025a). The main extensions include:

- **Multi-view EF estimation:** a new evaluation subset requiring both PLAX and A4C within each study, and a simple late-fusion strategy with improved accuracy (Section 5.2).
- **Expanded artifact release and usage documentation:** we further release Hugging Face model checkpoints, a Hugging Face Space demo, and a Google Colab notebook enabling convenient inference and evaluation with the released models and label files (Section 6). We also provide a concise artifact usage guide (Appendix B) to support reproducibility and ease of adoption.
- **Expanded methodological details for reproducibility:** we add implementation and hyperparameter details for the image-based bootstrap view classifier, the video view classifier, the A4C EF regressor, and the PLAX EF regressors in Appendix A.
- **Expanded clinical/technical context and literature:** a new Background section (Section 2) and a new Related Work section (Section 3) to better contextualize EF measurement, PLAX/A4C view characteristics, multi-view learning, and proxy supervision.

Organization. Section 2 summarizes EF measurement and the clinical role of A4C/PLAX views. Section 3 reviews prior work on EF estimation, PLAX modeling, multi-view echocardiography, and learning with scarce/noisy labels. Section 4 details our dataset generation pipeline and model training. Section 5 reports experimental results, and Section 6 summarizes released artifacts for reproducibility.

2. Background

This section provides clinical and technical background on EF measurement and the echocardiographic views most relevant to this work. We separate this material from Related Work to reduce redundancy and to clarify which aspects are established clinical practice versus new machine-learning contributions.

2.1 Left ventricular ejection fraction and its measurement

Left ventricular ejection fraction (EF) is defined as the fraction of blood ejected from the left ventricle (LV) during systole:

$$\mathrm{EF} = \frac{V_{\mathrm{ED}} - V_{\mathrm{ES}}}{V_{\mathrm{ED}}} \times 100\%, \tag{1}$$

where V_{ED} and V_{ES} are end-diastolic and end-systolic LV volumes, respectively. In guideline-recommended transthoracic echocardiography, LV volumes are commonly estimated via the (bi)plane method-of-disks (Simpson's rule), using endocardial contours traced in apical long-axis views (typically A4C and A2C) (Lang et al., 2015).

In contrast, PLAX-based EF estimation in clinical practice is frequently indirect and relies on geometric assumptions, for example by using linear LV internal dimensions (e.g., LVID) with analytic volume approximations such as the Teichholz formula (Teichholz et al., 1976). While such formulas can be fast and interpretable, they can be sensitive to image quality, view alignment, and regional wall

motion abnormalities that violate modeling assumptions. These limitations motivate end-to-end video-based learning approaches that attempt to map cine dynamics directly to EF without explicit intermediate measurements.

2.2 A4C and PLAX views

A standard echocardiography study contains multiple views that capture complementary anatomy. This work focuses on the apical four-chamber (A4C) and parasternal long-axis (PLAX) views:

- **A4C** provides a long-axis plane through both ventricles and atria, enabling volumetric assessment of the LV cavity and supporting guideline-recommended volume-based EF measurement (Lang et al., 2015). A schematic illustration of the A4C view is shown in Fig. 1a.
- **PLAX** provides a parasternal long-axis slice through the LV and left atrium, with prominent visualization of the LV outflow tract and mitral valve apparatus. PLAX is often robust for rapid qualitative LV function assessment and is commonly available in point-of-care workflows (Russell et al., 2022). A schematic illustration of the PLAX view is shown in Fig. 1b.

2.3 Public data resources and supervision gaps

A key challenge for PLAX EF modeling is that *public* supervision is misaligned with acquisition frequency: PLAX is common in practice, yet public PLAX datasets rarely provide EF labels. Table 1 summarizes representative public resources and the type of supervision they provide.

2.4 Study-level versus video-level targets

Clinical EF is typically reported at the *study* level, whereas ML datasets often provide labels per video clip. Real-world studies contain multiple clips per view and varying image quality. Averaging predictions across clips within a study can reduce variance and can better match the clinical reporting unit. This motivates our study-level evaluation protocol and our multi-view fusion strategy, both performed at the study level (Section 5).

3. Related Work

We review related work in four areas: (i) EF estimation from apical echocardiography (with emphasis on A4C), (ii) PLAX and non-apical EF estimation, (iii) multi-view learning in echocardiography, and (iv) learning with scarce and noisy supervision via proxy labels and report mining.

3.1 EF estimation from apical echocardiography

Most machine-learning work on echocardiographic EF estimation targets apical views because they support standardized volumetric measurements and are the dominant source of labeled EF in clinical reporting. EchoNet-Dynamic popularized end-to-end EF regression from A4C cine videos and demonstrated that spatiotemporal deep networks can achieve clinically competitive error when trained on curated EF labels (Ouyang et al., 2020). Typical pipelines map a fixed-length clip (or multiple sampled clips) to a scalar EF prediction and may incorporate temporal aggregation to stabilize predictions. Beyond direct regression, several approaches estimate EF through intermediate structure segmentation and volume computation, leveraging the method-of-disks or related geometric constructs; large-scale datasets such as CAMUS provide ED/ES annotations for segmentation and derived indices including EF (Leclerc et al., 2019).

Another practical concern is that apical views can be compromised by foreshortening, drop-out, and limited endocardial definition. Real-time systems have been proposed to jointly address EF estimation and view-quality failure modes (e.g., foreshortening detection) to improve reliability in deployment settings (Smistad et al., 2020). Collectively, these works show that A4C-based EF prediction is feasible and can be highly accurate *when* curated labels and consistent view definitions are available—a condition that does not hold for PLAX EF in public data.

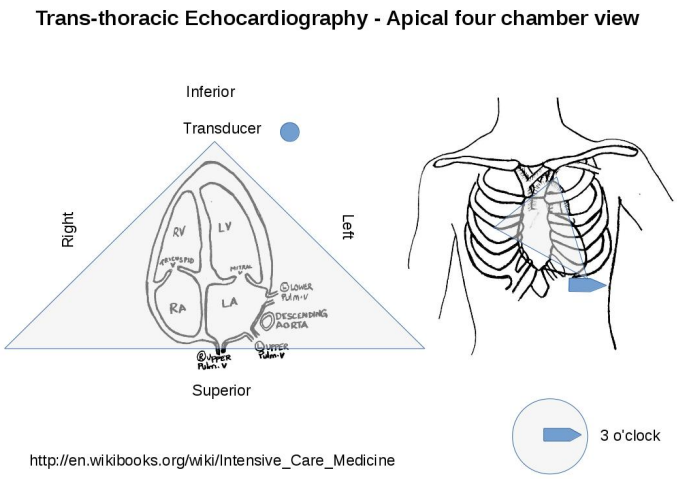
3.2 PLAX and non-apical EF estimation

PLAX is routinely acquired and frequently used in focused/point-of-care examinations, and clinical studies suggest that PLAX alone can be informative for identifying LV systolic dysfunction (Russell et al., 2022). However, PLAX is not commonly the primary view for EF reporting, resulting in a scarcity of PLAX–EF supervision.

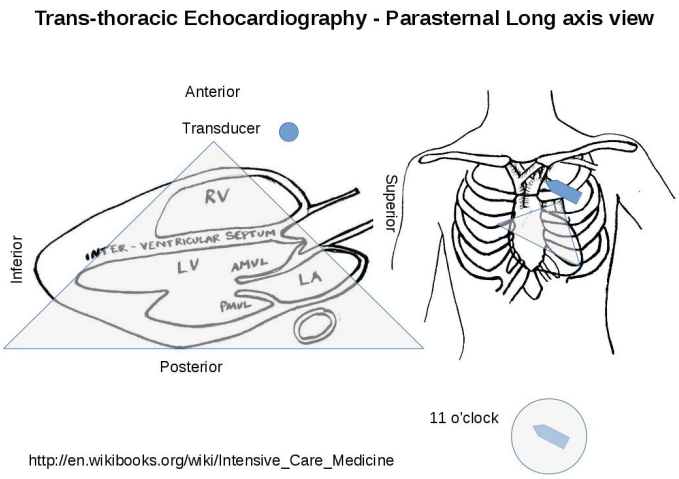
Existing PLAX EF efforts often follow one of two paradigms.

(i) Indirect measurement pipelines: EF is estimated from PLAX via intermediate surrogates such as linear LV dimensions (fractional shortening, LVID-based volume approximations) or landmark-based measurement extraction (Teichholz et al., 1976). Recent work proposed lightweight landmark detection on PLAX to estimate EF from sparse annotations, effectively learning to recover measurement-derived EF values (Goco et al., 2022). Such methods benefit from interpretability but introduce additional error sources, including landmark uncertainty and inter-observer measurement variability.

(ii) Direct PLAX EF prediction: Proprietary or closed systems have reported PLAX EF performance without sufficient disclosure for independent reproduction. For exam-



(a) Apical four-chamber (A4C) schematic.



(b) Parasternal long-axis (PLAX) schematic.

Figure 1: Standard long-axis echocardiographic views most relevant to this work. The schematics are reproduced from Wikimedia Commons under CC BY-SA 4.0 (au, 2016a,b).

Table 1: Representative public echocardiography resources used or referenced in this work and their supervision.

Resource	View(s)	Supervision
EchoNet-Dynamic (Ouyang et al., 2020)	A4C	EF, volumes
EchoNet-LVH (Duffy et al., 2022)	PLAX	LVH-related labels
CAMUS (Leclerc et al., 2019)	A4C, A2C	Segmentation, volumes, EF
MIMIC-IV-Echo (Gow et al., 2023; Goldberger et al., 2000)	multi-view	none (raw video)
MIMIC-IV-Note (Johnson et al., 2023)	text	discharge notes
TMED-2 (Huang et al., 2022)	multi-view (images)	view labels

ple, ExoAI reported PLAX-specific error in a retrospective evaluation, but training data and algorithmic details were not fully available to the community (Vega et al., 2024). More recently, point-of-care studies have explored PLAX-based EF estimation via motion/wall tracking and related approaches (Vega et al., 2025). Overall, the field lacks a widely reproducible, public PLAX–EF benchmark with runnable artifacts, limiting systematic comparisons across methods.

3.3 Multi-view learning in echocardiography

Echocardiography is inherently multi-view: a single study typically includes multiple standardized planes (parasternal and apical views) that emphasize different anatomy and motion patterns. Multi-view learning has therefore been explored across tasks including view classification, segmentation, disease classification, and quantitative function estimation. Fusion strategies range from *late fusion* (combining independently trained view-specific predictors) to learned fusion via attention or global-local aggregation modules (Zheng et al., 2023).

For EF-related tasks, multi-view modeling is appealing because different views fail for different reasons (e.g., apical foreshortening versus parasternal drop-out). Prior clinical ML systems have adapted EF estimation to work from a subset of commonly acquired point-of-care views (including PLAX and A4C), and to leverage whichever views are available (Asch et al., 2019, 2021). Recent architectures explicitly ingest multiple standard views via multi-stream encoders and transformer-based fusion, often combining appearance and motion cues (e.g., optical flow) to improve robustness (Alvén et al., 2025). Additionally, transformer-based multi-view encoders have been developed to aggregate view-level representations into study-level embeddings for downstream tasks, highlighting the broader trend toward study-level, multi-view representations (Tohyama et al., 2025). Our work contributes to this line by providing a reproducible A4C–PLAX multi-view EF test subset and demonstrating that even simple unweighted late fusion yields measurable gains when both views are present.

3.4 Learning with scarce and noisy labels: proxy supervision and report mining

Label scarcity and label noise are pervasive in clinical ML. When the desired supervision is not routinely recorded for a given view or modality, a common approach is to use *proxy labels* or *pseudo-labels* from a teacher model (self-training), enabling large-scale training despite limited ground truth (Lee, 2013; Xie et al., 2020). Weak supervision frameworks further formalize the combination of multiple noisy sources (heuristics, rules, imperfect models) to create training labels at scale (Ratner et al., 2016).

In echocardiography, EF is often present in unstructured reports rather than machine-readable fields. Prior work has extracted EF values from clinical text using NLP pipelines, enabling retrospective cohort construction from notes (Kim et al., 2017). More recently, large language models have been used as flexible extractors of structured information from heterogeneous clinical narratives; in our pipeline we use GPT-4-based extraction with subsequent validation against video-model predictions to improve reliability (OpenAI, 2024). Combining proxy-supervised training with an *independent* note-derived ground-truth evaluation cohort allows us to (i) learn PLAX EF models at scale and (ii) transparently quantify performance without circularly evaluating on the teacher-generated labels.

4. Dataset Generation and Model Training

The key step of our methodology—and the major challenge—was creating a labeled dataset for PLAX from available data, despite the scarcity of publicly available PLAX-specific labels. Because PLAX images and EF labels are not routinely paired in existing repositories, we developed novel techniques to both identify PLAX views and assign approximate EF values, effectively circumventing the lack of direct ground-truth annotations.

Two major datasets from PhysioNet(Goldberger et al., 2000) were utilized for this study:

- **MIMIC-IV-Echo**(Gow et al., 2023): Contains approximately 500k echocardiographic videos without view type labels.
- **MIMIC-IV-Note**(Johnson et al., 2023): Includes around 331k discharge notes, of which only a subset contains EF values in unstructured text.

The goal was to extract PLAX view videos with EF data and use them to train a machine learning model for EF prediction. This task faced two main challenges: first, the MIMIC-IV-Echo dataset lacked labels for echocardiographic view types, requiring the development of a video view classifier to identify PLAX views; second, the MIMIC-IV-Echo and MIMIC-IV-Note datasets were not directly linked, making it difficult to associate discharge notes with corresponding echocardiographic studies. Even after applying time-based correlations, the number of valid note-study pairs remained insufficient for training a robust PLAX model.

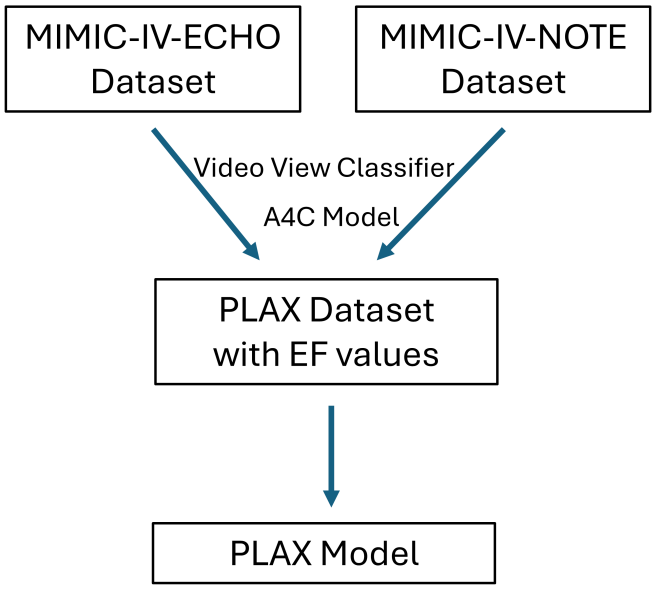


Figure 2: High-Level PLAX EF Prediction Pipeline.

Consequently, we needed to leverage most studies in the MIMIC-IV-Echo dataset to generate training data for PLAX EF prediction. Specifically, we first trained a video view classifier to identify A4C and PLAX views within the dataset. Next, we trained an A4C model using a publicly available dataset to estimate EF values. These EF predictions were then applied as proxy labels for PLAX videos within the same study, enabling the development of a PLAX-specific model. A high-level overview is illustrated in Fig. 2.

4.1 Video View Classifier Training

A classifier capable of distinguishing echocardiographic views into A4C, PLAX, and "OTHER," was essential for accurately selecting A4C and PLAX videos from the MIMIC-IV-Echo dataset.

We first leveraged the **TMED-2** dataset (Huang et al., 2022), which contains approximately 25,000 labeled echocardiographic images spanning a range of views, including A4C, PLAX, apical two-chamber (A2C), parasternal short-axis (PSAX), and a combined category of miscellaneous views labeled as "A4C/A2C/OTHER." Using these labeled images, we trained a ResNet-34 image classifier.

The image classifier was applied frame-by-frame to MIMIC-IV-Echo videos, and aggregated predictions were used to assign video-level classifications. The model produced log-softmax outputs, which were converted to probabilities via the exponential transform, and the class with

the highest score was taken as the predicted view.

However, the MIMIC-IV-Echo dataset does not provide ground-truth labels for the videos, preventing direct measurement of the image classifier's accuracy on this dataset.

To evaluate performance on video data in the absence of ground-truth annotations, two independent reviewers manually assessed 300 randomly selected test videos. The evaluation results are summarized below:

- For 100 videos predicted as A4C, two reviewers identified 91/87 as correct.
- For 100 videos predicted as PLAX, two reviewers identified 78/72 as correct.
- For 100 videos not predicted as A4C/PLAX, two reviewers identified 94/98 as correct.

These findings indicated that while the image classifier was highly reliable in identifying "OTHER" views, its accuracy for A4C and PLAX was insufficient for downstream use, motivating the development of a more robust video-based classifier. To train such a classifier, we required videos labeled as A4C, PLAX, and "OTHER" views. This process is outlined in the flowchart in Fig. 3.

For A4C and PLAX training data, we utilized the following two EchoNet datasets. EchoNet is a publicly available echocardiographic dataset designed for ML research. Unlike the MIMIC datasets, EchoNet datasets are highly curated collections of echocardiographic videos with specific annotations for view types and clinical parameters.

- **EchoNet-Dynamic**(Ouyang et al., 2020): Approximately 10,000 videos labeled as A4C with EF values.
- **EchoNet-LVH**(Duffy et al., 2022): Approximately 12,000 videos labeled as PLAX without EF values.

Constructing an "OTHER" video dataset was more challenging, as it required broad representation of alternative views. To address this, we leveraged the image classifier (which showed reliable performance for "OTHER") to extract videos from MIMIC-IV-Echo. Specifically, we selected:

- 4,212 A2C videos (exp-transformed score > 0.6),
- 4,015 PSAX videos (exp-transformed score > 0.6),
- 6,000 videos labeled as "A4C/A2C/OTHER" (exp-transformed score > 0.9).

The exp-transformed scores were manually set to balance the number of videos across different categories, ensuring adequate representation for training and minimizing contamination:

- First, for better model performance, it's desirable to balance A4C, PLAX, and "OTHER" categories, targeting "OTHER" dataset sizes similar to EchoNet-Dynamic (10,030 A4C) and EchoNet-LVH (12,000 PLAX).
- Second, due to the limitations of the TMED dataset, the "OTHER" category was constructed using A2C, PSAX, and "A4C/A2C/OTHER" labels. Within "A4C/A2C/OTHER", manual verification showed that increasing the exp-transformed score threshold reduced A4C/A2C contamination. Thus, we set 0.9 to ensure diversity while minimizing A4C inclusion.
- Third, to balance A2C and PSAX within "OTHER", while maintaining overall dataset proportionality between A4C, PLAX and "OTHER", we set a 0.6 threshold for both A2C and PSAX. This approach allowed us to create a comprehensive "OTHER" dataset while maintaining diversity in the video views.

Finally, the overall distribution of training data for the video view classifier was as follows:

- **A4C:** 10,030 videos, sourced entirely from the EchoNet-Dynamic dataset.
- **PLAX:** 12,000 videos, sourced entirely from the EchoNet-LVH dataset.
- **"OTHER":** 14,227 videos, ML-classified from the MIMIC-IV-Echo dataset.

For the final video view classifier, we utilized a pretrained X3D-s model (Feichtenhofer, 2020) on this dataset. Compared with the ResNet-34 image classifier, the video-based model demonstrated substantially higher accuracy in distinguishing A4C and PLAX views, ensuring higher-quality data for subsequent analyses. This video-based classifier formed the backbone of our video view classification pipeline, achieving robust performance and enabling accurate selection of PLAX and A4C videos for downstream tasks.

Additional implementation details for the view-classification models (image-based bootstrapping on TMED-2, video-model fine-tuning, clip sampling, data augmentation, optimization, and compute setup) are provided in Appendix A.1.

4.2 A4C Model Training

We train an A4C EF regression model on **EchoNet-Dynamic**, which contains 10,030 A4C cine videos with expert-reported EF labels. Following Ouyang et al. (2020), we implement a spatiotemporal **R(2+1)D** network (TorchVision `r2plus1d_18`) initialized from a public

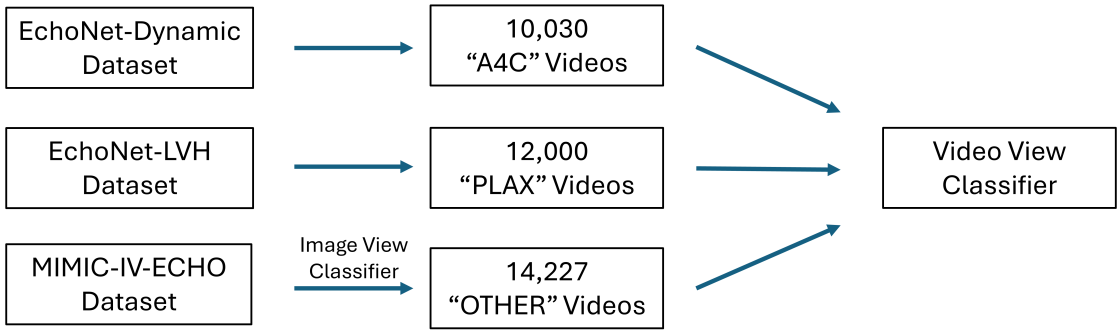


Figure 3: Multi-Dataset Strategy for Video View Classification.

video-pretrained checkpoint, replacing the final classification layer with a single-neuron regression head. Training and model selection follow the official EchoNet-Dynamic `TRAIN/VAL/TEST` split, with mean squared error (MSE) as the regression loss.

The resulting model achieves an MAE of **4.37%** on the test split, closely matching the 4.1% MAE reported in Ouyang et al. (2020). Full implementation details (clip sampling, preprocessing/augmentation, optimization, and checkpoint selection) are provided in Appendix A.2. This trained A4C predictor is subsequently used as a teacher model to generate study-level proxy EF labels for PLAX training (Section 4.5).

4.3 Ground Truth Data Generation

A ground truth dataset was essential for evaluating the true error of the PLAX EF model. Because the PLAX EF model was trained on EF values indirectly generated by the A4C model, its performance inherently included the compounded error of A4C predictions. To measure the error of the PLAX EF model independently, we constructed a ground truth dataset with directly validated EF values. An overview of this process is shown in Fig. 4.

No direct mapping exists between studies in the MIMIC-IV-Echo and MIMIC-IV-Note datasets. To bridge this gap, we performed a time-based correlation between the two datasets using the same patient ID. This approach identified 2,560 note-study pairs where echocardiography studies and clinical notes were recorded within a 60-day window.

Given the unstructured and messy text format of the notes, we utilized the GPT-4 API(OpenAI, 2024) to extract EF values and other relevant information. The API was instrumental in parsing and cleaning the free-text notes to retrieve meaningful clinical data, which was also publicly available (Section 6). This automated parsing and cleaning process yielded 921 valid note-study pairs within the 60-day window with a valid EF value.

To assess whether these note-derived EF values are reliable rather than artifacts of a single extractor, we repeated the extraction with two additional, independent and more recent large language models (GPT-5.5 and Claude Opus 4.8) using an identical prompt. Across the notes for which all three models returned a numeric EF, the three models agreed exactly in 99.5% of cases, with pairwise Pearson correlations $r \approx 1.0$. These results support the reproducibility of the note-derived EF extraction and reduce the likelihood that the evaluation labels are driven by model-specific extraction artifacts. For consistency with the originally released benchmark and to avoid changing the evaluation protocol, we retained the GPT-4-extracted EF values as the final labels.

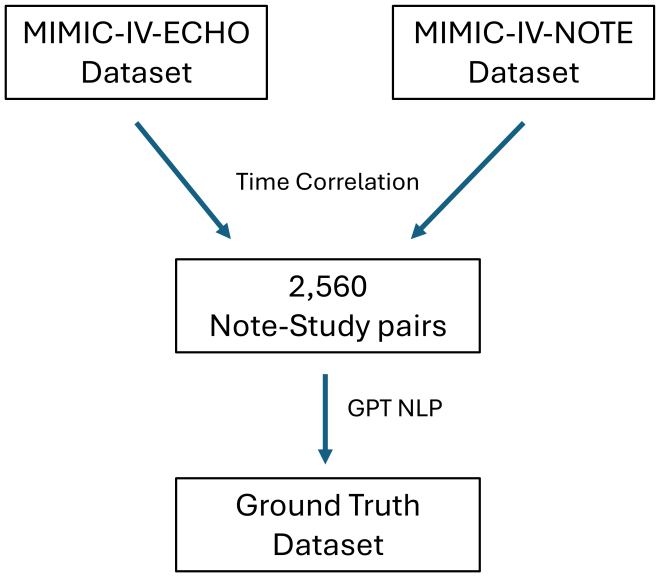


Figure 4: Ground Truth Dataset Generation Flow Chart.

Next, we applied the video view classifier from Section 4.1 to identify studies containing at least one candidate A4C clip and one candidate PLAX clip. After this view-based filtering, 902 note–study pairs remained. We then enforced a confidence threshold by requiring an exp-transformed score greater than 0.5 for both the A4C and PLAX predictions, yielding 848 high-confidence pairs.

To validate the note-extracted EF values, we ran the A4C model described in Section 4.2 on the A4C videos for the studies and compared the predictions with the note labels, obtaining a study-level MAE of 7.82%. To further improve accuracy, we restricted the correlation to a 1-day window between notes and studies. This yielded 295 high-confidence note–study pairs, each containing at least one PLAX and one A4C video. In addition, videos were screened to ensure a duration greater than one second and to exclude studies containing color Doppler imaging. After this filter, the final test set included 290 studies with 1,017 A4C videos and 295 studies with 1,320 PLAX videos (note that the study counts overlap).

On this high-confidence cohort, the A4C model achieved a MAE of 6.95% when compared against note-extracted labels. This closely matches the reported out-of-sample performance of 6.0% in (Ouyang et al., 2020), providing evidence that the note-derived EF values are reliable. These 295 studies (comprising 1,017 A4C videos and 1,320 PLAX videos) formed our ground truth test set, and their labels are publicly available (Section 6).

Although this ground truth dataset was not large enough to train the PLAX EF model, it serves as an independent benchmark. The error measured on this set provides a robust evaluation of the PLAX EF model that does not depend on proxy EF values generated by the A4C model.

4.4 View Classifier Fine-Tuning

The video view classifier is essential to our pipepline. To further enhance the performance of our video view classifier (X3D model), we leveraged the 902 valid note-study pairs identified within the 60-day window as described in Section 4.3. These pairs were used to further refine the classifier's ability to identify A4C views.

First, we applied the view classifier to extract A4C videos from studies, identifying a total of 4,131 videos within the 60-day window. Next, we ran our A4C model on these identified A4C videos, producing the error distribution shown in Fig. 5.

Upon reviewing the 10 videos with the highest errors, we found that their views were either not true A4C views or partial A4C views. This observation demonstrated that the errors of the A4C model could be effectively utilized to fine-tune the video view classifier.

To address this, we fine-tuned the video view classifier using a subset of videos with extreme errors:

- Videos with MAE $< 3\%$ were labeled as A4C.
- Videos with MAE $> 20\%$ were labeled as "OTHER."

This subset comprised 1,346 videos, which were split equally into training and test sets for fine-tuning.

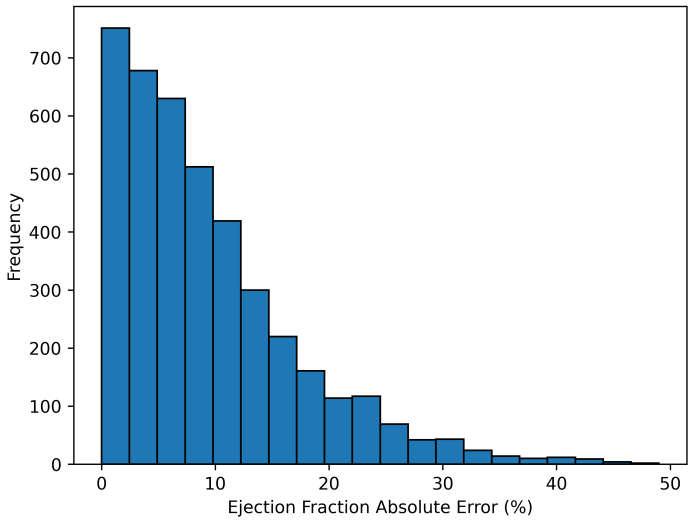


Figure 5: Error distribution of the A4C model within 60-day window.

The X3D model was fine-tuned using the labeled training set, with the objective of improving its ability to distinguish between A4C and "OTHER" views. After fine-tuning, the test set was regenerated using the fine-tuned classifier, and the MAE was recalculated using the A4C model. The MAE was reduced from 6.83% to 5.14%, demonstrating a significant improvement in classifier accuracy.

This fine-tuning process enabled the classifier to more accurately identify A4C views, reducing misclassifications and ensuring higher-quality input for downstream tasks.

4.5 PLAX Dataset Generation

We applied the video view classifiers, both before and after fine-tuning, to the MIMIC-IV-Echo dataset, where the patients appearing in the ground truth dataset were excluded. An overview of this process is shown in Fig. 6.

The label distribution of videos before fine-tuning is shown in Fig. 7a, while the distribution after fine-tuning is shown in Fig. 7b. After fine-tuning, the classifier identified A4C views more strictly, reducing the number of videos labeled as A4C by approximately 15,000, resulting in a total of around 20,000 videos. Conversely, the number of videos classified as PLAX views increased significantly.

To ensure maximum accuracy, we selected PLAX views from the pre-fine-tuning classifier and A4C views from the post-fine-tuning classifier, eliminating any overlapping videos. Most studies contained both an A4C video and a PLAX video. For A4C videos in each study, we used the A4C model described in Section 4.2 to generate EF values, and then averaged those values to serve as the EF label for the PLAX videos in the study.

The final PLAX training dataset consisted of 25,532 videos, comprising 4,822 studies (80% training, 20% valida-

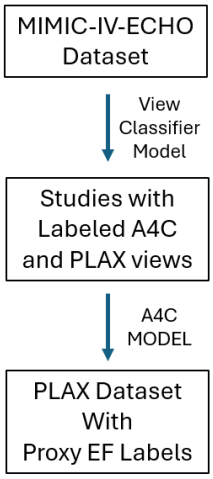


Figure 6: Overview of the PLAX Dataset Construction Pipeline Using Proxy Supervision.

tion). Different videos within the same study were assigned to the same split for a clean validation set. Labels are publicly available (Section 6).

For testing, we used the ground truth dataset described in Section 4.3. Using the pre-fine-tuned view classifier, which is stricter in identifying PLAX views, we extracted 1,320 PLAX videos from the 295 studies. To ensure strict data separation, patients included in the test set were excluded from both the training and validation sets.

With the proxy-labeled training set and the held-out ground-truth cohort in place, we next train PLAX EF models and report quantitative results.

5. Experimental results and analyses

Building on the proxy-labeled PLAX training set constructed in Section 4.5 and the independent note-derived ground-truth cohort described in Section 4.3, we now evaluate PLAX-based EF prediction and quantify the benefits of multi-view integration. Unless otherwise stated, we report *study-level* performance by averaging video-level predictions across all clips of a given view within each study, which better matches the clinical unit of EF reporting (Section 2). We first present single-view PLAX results, and then analyze a simple late-fusion strategy that combines A4C and PLAX predictions on studies containing both views.

5.1 PLAX Model Training and Results

We trained two R(2+1)D models with different configurations based on the PLAX training dataset mentioned in section 4.5. The training details and configurations are summarized in Table 2.

Table 2: Training configurations for PLAX EF prediction models.

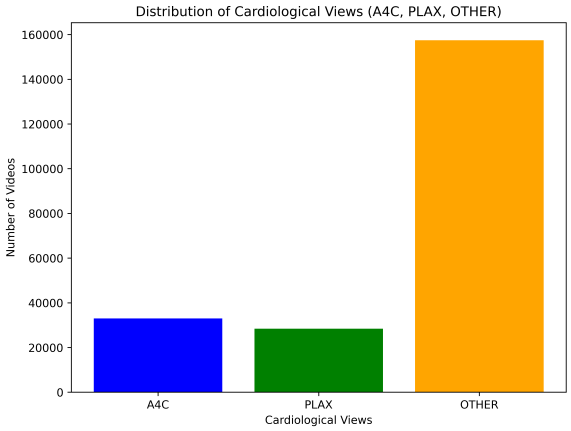
Model	Batch Size	MAE	Correlation
R(2+1)D	16	6.93%	0.659
R(2+1)D	32	6.89%	0.656

Both models were trained for 100 epochs using a learning rate (LR) scheduler with an initial LR of 0.001, a patience of 5 epochs and a reduction factor of 0.1. Input clips were spatially standardized using padding and random cropping to a fixed resolution. Standard preprocessing, including padding and random cropping, was applied. For each configuration, the checkpoint with the lowest validation error was selected as the final model. Additional training and implementation details for the PLAX models are provided in Appendix A.3.

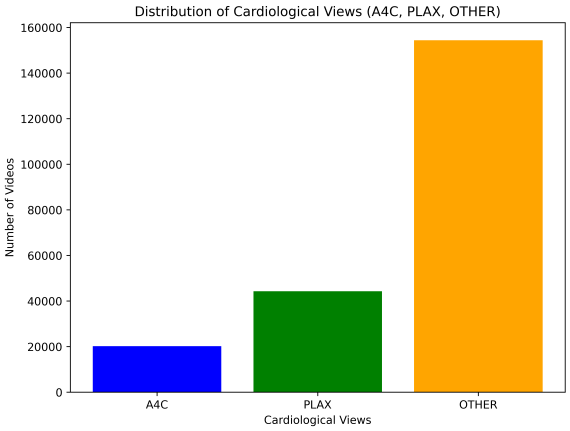
At test time, study-level predictions were obtained by averaging video-level predictions, and MAE was computed per study. The final EF estimate was produced by taking an unweighted average of the two model outputs on the ground truth dataset described in Section 4.3, yielding an MAE of **6.86%**.

To further assess the agreement between predicted and true EF values, we computed the Pearson correlation coefficient of 0.670, indicating a reliably positive correlation between our model's predictions and the ground truth. The scatter plot (Fig. 8) demonstrates that predictions generally follow the line of identity, though variance increases for higher EF values. Additionally, we present a Bland-Altman plot (Fig. 9) to analyze systematic bias and agreement limits. The mean difference (bias) is -0.62%, suggesting no significant systematic offset in predictions. The upper and lower limits of agreement (LoA) are 16.87% and -18.10%, respectively, indicating some variability in the prediction errors. Notably, the plot shows increased dispersion at higher EF values, which aligns with previous observations in EF estimation models.

One potential explanation for this variance and the moderate correlation is the use of proxy labels derived from an A4C-based EF model, which can introduce compounded errors. Additionally, since the A4C model was not originally trained on MIMIC datasets, its predictions can introduce some domain shift. However, despite these challenges, our work provides, to the best of our knowledge, the first *publicly reproducible* PLAX EF benchmark with disclosed methodology and runnable inference artifacts. On this cohort, our PLAX model achieves lower MAE than previously reported PLAX EF results we are aware of, including proprietary systems with limited disclosure and indirect measurement-based pipelines. Prior works either lack methodological transparency (e.g., ExoAI with an MAE of 7.29%) (Vega et al., 2024) or rely on indirect EF estimation methods such



(a) Label distribution before fine-tuning.



(b) Label distribution after fine-tuning.

Figure 7: Comparison of label distributions before and after fine-tuning the video view classifier.

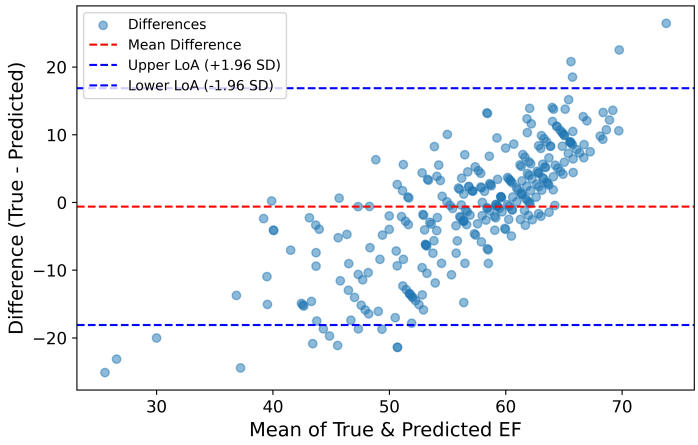


Figure 8: Scatter Plot for PLAX EF prediction. The dashed red line indicates the line of identity.

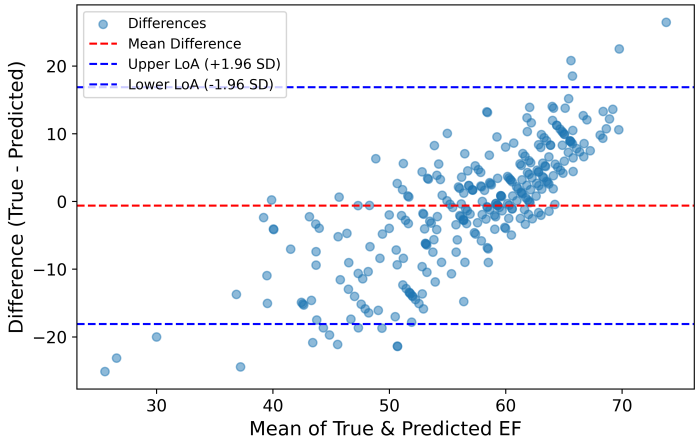


Figure 9: Bland-Altman Plot for PLAX EF prediction.

as LVID measurement (MAE 8.45%) (Goco et al., 2022), which introduce inter-observer variability. By directly predicting EF from full PLAX cine videos, our approach avoids these manual dependencies while setting a reproducible standard for future research.

5.2 Aggregating A4C and PLAX Views

Because most echocardiographic studies contain both A4C and PLAX views, we investigated whether integrating information across views could further improve EF estimation. The intuition is that the two perspectives capture complementary anatomical details, and combining them may reduce variance and correct for view-specific biases.

We used the ground truth dataset described in Section 4.3, which includes 290 studies with 1,017 A4C videos and 295 studies with 1,320 PLAX videos (with partial overlap between the study sets). To construct a multi-view test set, we retained only studies containing at least one A4C video and one PLAX video. This yielded 284 studies comprising 1,000 A4C videos and 1,275 PLAX videos. These studies formed our multi-view test set, which is publicly available (Section 6).

We implemented a late-fusion strategy at the study level by taking an unweighted average of the A4C model prediction and the PLAX ensemble prediction. This simple approach required no additional training and preserved the independence of the single-view models.

As summarized in Table 3, the fusion outperformed both single-view baselines, reducing MAE to 6.37% and increasing Pearson correlation to 0.709. This demonstrates that integrating complementary views yields a tangible gain in study-level EF prediction accuracy.

The fusion gain is incremental: the two views share overlapping information about LV systolic function, and the PLAX model is trained from A4C-derived proxy labels. Still, the improvement shows that simple study-level late fusion

can add signal beyond either single-view prediction alone.

Table 3: Study-level performance comparison of single-view and multi-view EF prediction. *Evaluated on the 284-study multi-view subset (studies containing both A4C and PLAX); results are not directly comparable to Section 5.1, which uses the full PLAX-only ground-truth cohort containing 295 studies.*

Model	MAE	Correlation
A4C-only	7.01%	0.657
PLAX-only (ensemble)	6.77%	0.676
A4C + PLAX (fusion)	**6.37%**	**0.709**

Agreement and calibration. The scatter plot (Fig. 10) shows that fusion predictions align closely with the line of identity, with visibly reduced dispersion compared to single-view models. The Bland-Altman analysis (Fig. 11) confirms this improvement: the mean bias was –0.08%, with limits of agreement at +15.93% and –16.08%. Compared to the PLAX-only results, both the bias and error spread were reduced, indicating better calibration and more consistent predictions across the EF range.

Single-view comparison. One thing worth noting is the comparison between the two single views. Although the PLAX-only ensemble reports a numerically lower MAE than the A4C-only model (6.77% vs. 7.01%) on this subset, this difference is not statistically significant: on the 284 paired studies, a study-level paired comparison of absolute errors yields no significant difference (Wilcoxon signed-rank $p = 0.66$; paired t-test $p = 0.48$; $\Delta\text{MAE} = 0.24\%$, 95% bootstrap CI $[-0.40, 0.88]$). The two single views are therefore statistically comparable rather than one being superior; in particular, the PLAX student matches but does not exceed its A4C teacher, consistent with proxy supervision introducing no view-level accuracy loss.

Proportional bias. The Bland–Altman plot (Fig. 11) also indicates a proportional bias, with overestimation at lower EF values and underestimation at higher EF values. To quantify this, we regressed the per-study prediction differences (predicted minus reference EF) on the mean EF for each setting. All three views exhibit a statistically significant negative slope:

- **A4C:** slope -0.22 (95% CI $[-0.33, -0.12]$, $p < 0.001$);
- **PLAX:** slope -0.88 (95% CI $[-0.97, -0.80]$, $p < 0.001$);
- **Fusion:** slope -0.58 (95% CI $[-0.67, -0.49]$, $p < 0.001$).

The pattern across views is informative as to cause. The A4C model, trained on expert-annotated EchoNet-Dynamic EF labels, still exhibits a significant slope, indicating that MSE-driven shrinkage toward the population mean is present even under clean supervision. The PLAX model, trained on A4C-derived proxy labels, shows the steepest slope, consistent with proxy supervision and note-derived label noise adding further shrinkage on top of this baseline; the fusion result falls between the two, as expected from averaging the two views.

Inter-view disagreement. As a further exploratory analysis, we tested whether the disagreement between the two single-view predictions, $|\text{EF}_{\text{A4C}} - \text{EF}_{\text{PLAX}}|$, is associated with the final fusion error. The association is weak and small in magnitude: the regression slope is ≈ 0.08 EF points of error per EF point of disagreement (Pearson $r = 0.07$, $p = 0.21$; Spearman $r = 0.13$, $p = 0.04$), and a quartile analysis shows no consistently monotonic trend. Inter-view disagreement may therefore provide at most a weak reliability signal and is not, on this dataset, sufficient to serve as a standalone uncertainty measure.

Alternative fusion rules. We further examined whether a more elaborate fusion rule would improve on the unweighted average. On the 284 common studies, a learned convex weight $w, \text{EF}{*}\text{A4C}{+}(1{-}w), \text{EF}{*}\text{PLAX}$ selected $w \approx 0.44$, nearly recovering the equal-weight average, and did not reduce study-level MAE (5-fold cross-validated 6.38% vs. 6.37%; Wilcoxon signed-rank $p = 0.38$). A regularized linear fusion (ridge regression over both view predictions with an intercept) likewise showed no significant difference (6.34% vs. 6.37%; $p = 0.87$). Learned fusion rules can in principle calibrate view-specific biases and assign data-driven weights, but they also introduce additional fitted parameters and greater overfitting risk in this limited paired-data setting. In contrast, the unweighted average is less flexible but parameter-free and preserves the independence of the single-view models.

Summary. Together, these analyses highlight the complementary value of integrating multiple echocardiographic views. The simple study-level fusion improved over both single-view baselines without additional video-model training. Beyond establishing a benchmark for PLAX EF prediction, our results suggest that multi-view integration can serve as a practical strategy for more dependable EF assessment in real-world echocardiography workflows.

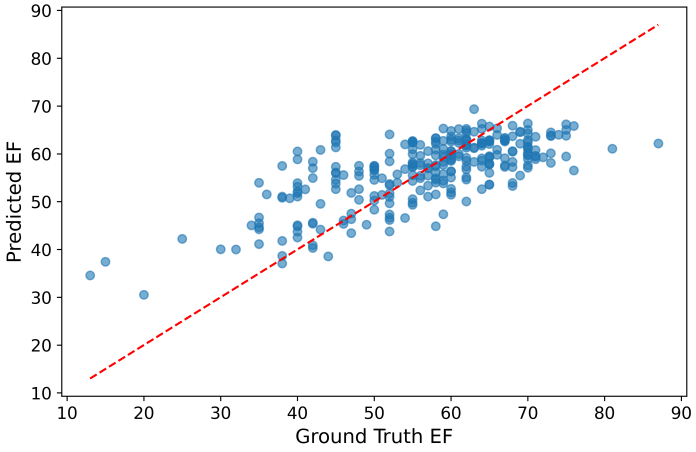


Figure 10: Scatter plot for A4C + PLAX fusion EF predictions against ground truth. The dashed red line indicates the line of identity.

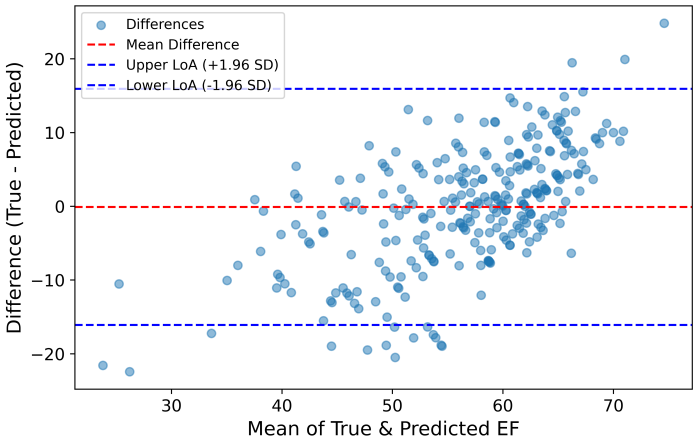


Figure 11: Bland-Altman plot for A4C + PLAX fusion EF predictions.

6. Artifacts and Reproducibility

To support reproducibility, transparency, and further research, we release a complete set of artifacts that enable reconstruction of the evaluation cohorts via released label files (given authorized access to MIMIC-IV resources), and convenient inference with pretrained A4C and PLAX EF models. A concise usage guide is provided in Appendix B.

- **GitHub repository.** We provide dataset label files, instructions for locating the corresponding videos within MIMIC-IV-Echo under the data-use agreement, and the GPT prompts used for EF extraction from clinical notes.

  `https://github.com/Jeffrey4899/PLAX_EF_Labels_202509`

- **Hugging Face Space.** A browser-based interactive demo for running inference with the released EF models (A4C/PLAX/multi-view fusion) on sample videos or user-uploaded inputs (research and education use only).

  `https://huggingface.co/spaces/Jeff4899/202509_PLAX_EF_Demo`

- **Trained models on Hugging Face.** Inference-ready models for both PLAX EF and A4C EF prediction are released for direct download and integration into existing pipelines.

  `https://huggingface.co/Jeff4899/202509_PLAX_EF`
  `https://huggingface.co/Jeff4899/202509_A4C_EF`

- **Google Colab demo.** An interactive notebook that enables users to evaluate the released models on sample echocardiography videos or their own uploads, with full access to and modification of the underlying code.

  `https://colab.research.google.com/drive/1E2IWrfpBIKI4cBoBTCn3OLwEK9o3GTMM?usp=sharing`

Together, these artifacts are intended not only to facilitate reproduction of the reported results, but also to serve as a foundation for future methodological and clinical extensions.

7. Conclusion

We introduced, to the best of our knowledge, the first publicly available label resource enabling EF estimation from parasternal long-axis (PLAX) cine echocardiography, together with a reproducible pipeline that learns from scarce labels by mining views, extracting EF from clinical notes, and training PLAX models with proxy supervision. On an independent note-derived ground-truth cohort, our PLAX ensemble achieved a study-level MAE of 6.86%. We further

showed that simple study-level late fusion of A4C and PLAX predictions improves accuracy to 6.37% MAE (correlation 0.709) on studies containing both views, highlighting the practical value of multi-view integration for EF estimation.

We release label files, pretrained model weights, and runnable inference demos to facilitate reproduction of the reported evaluation (given authorized access to the underlying datasets) and to support downstream research.

Clinical utility. The clinical motivation for this work is based on two practical considerations. First, PLAX is a routinely acquired echocardiographic view, while EF estimation from apical views depends on sufficient apical image quality and can be affected by issues such as foreshortening (Lang et al., 2015). Therefore, PLAX-based EF estimation may extend automated assessment to scans where standard apical windows are unavailable, incomplete, or suboptimal. Second, PLAX remains under-labeled for EF estimation in public resources. For example, EchoNet-LVH provides a large public PLAX video dataset, but its labels focus on chamber size and wall-thickness measurements rather than EF (Duffy et al., 2022). To our knowledge, no public PLAX–EF benchmark currently exists. Therefore, this work provides a reproducible PLAX–EF benchmark, proxy-supervision pipeline, and baseline models for a clinically relevant but under-labeled view.

Limitations. Our PLAX models are trained with proxy supervision derived from an A4C teacher, so performance is inherently affected by teacher noise and may be bounded by the A4C predictor under domain shift. In addition, view classification errors can propagate into both cohort construction and proxy labeling. Our ground-truth evaluation relies on EF values extracted from clinical notes with a strict time-window heuristic, which may still introduce residual label noise when EF reporting and imaging are not perfectly aligned. All released artifacts are intended for research and education use only; further external validation on independent clinical datasets would be required before considering any clinical deployment.

Future work. To further improve and validate our model, we are initiating a collaboration with a leading heart hospital to leverage their clinical datasets for refining label accuracy and enhancing generalizability. Future work will focus on incorporating expert-annotated PLAX EF values, for which a dataset is currently being secured. Additionally, we plan to evaluate clinical applicability through external validation and potential clinical trials, ensuring real-world effectiveness in echocardiographic workflows.

Taken together, our results suggest that multi-view integration offers a practical and effective way to improve upon A4C-only prediction, while also demonstrating that PLAX views alone can provide meaningful EF estimates when A4C views are unavailable or suboptimal.

Acknowledgments

This work was supported by the Center for Sensing to Intelligence (S2I) at Caltech and by Charles R. Trimble.

The authors thank the contributors of the MIMIC-IV-Echo, EchoNet, and TMED-2 datasets for making their data publicly available. We also thank the MIDL 2025 reviewers for their constructive feedback, which helped improve the clarity and scope of this extended version.

Ethical Standards

This study used only publicly available, de-identified datasets (MIMIC-IV-Echo, MIMIC-IV-Note, and EchoNet) released under their respective data use agreements. Institutional review board approval was not required because no new data were collected and no human subjects were directly involved in this study. All analyses were conducted in compliance with the ethical standards of the data providers and applicable regulations.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Data availability

All artifacts required to reproduce the results of this study (given authorized access to MIMIC-IV resources) are publicly available at the following URLs:

- **GitHub repository (labels, prompts, instructions):** `https://github.com/Jeffrey4899/PLAX_EF_Labels_202509`
- **Hugging Face Space (interactive demo):** `https://huggingface.co/spaces/Jeff4899/202509_PLAX_EF_Demo`
- **Hugging Face models (PLAX EF):** `https://huggingface.co/Jeff4899/202509_PLAX_EF`
- **Hugging Face models (A4C EF):** `https://huggingface.co/Jeff4899/202509_A4C_EF`
- **Google Colab demo:** `https://colab.research.google.com/drive/1E2IWrfpBIKI4cBoBTCn3OLwEK9o3GTMM?usp=sharing`

Due to data use agreements, raw MIMIC-IV-Echo videos are not redistributed. The released label files provide identifiers and instructions for locating the corresponding samples under authorized access.

References

Johannes Alvén, Eva Hagberg, David Hagerman, Ola Hjelmgren, et al. A deep multi-stream model for robust left ventricular ejection fraction prediction in two-dimensional echocardiography. *Scientific Reports*, 15:8734, 2025. . URL https://www.nature.com/articles/s41598-025-18766-3.

Frances M. Asch et al. Automated assessment of left ventricular ejection fraction and longitudinal strain by deep learning. *Circulation: Cardiovascular Imaging*, 12 (9):e009303, 2019. . URL https://doi.org/10.1161/CIRCIMAGING.119.009303.

Frances M. Asch et al. Deep learning-based automated echocardiographic assessment of left ventricular ejection fraction in the point-of-care setting. *Circulation: Cardiovascular Imaging*, 14(6):e012293, 2021. . URL https://doi.org/10.1161/CIRCIMAGING.120.012293.

Balaji.md au. Echoapicalfourchamber.jpg: Schematic representation of apical four chamber view in transthoracic echocardiography. Wikimedia Commons, 2016a. URL https://commons.wikimedia.org/wiki/File:Echoapicalfourchamber.jpg. CC BY-SA 4.0.

Balaji.md au. Echoparasternallongaxis.jpg: Schematic representation of parasternal long axis view in transthoracic echocardiography. Wikimedia Commons, 2016b. URL https://commons.wikimedia.org/wiki/File:Echoparasternallongaxis.jpg. CC BY-SA 4.0.

Grant Duffy, Paul P. Cheng, Neal Yuan, Bryan He, Alan C. Kwan, Matthew J. Shun-Shin, Kevin M. Alexander, Joseph Ebinger, Matthew P. Lungren, Florian Rader, David H. Liang, Ingela Schnittger, Euan A. Ashley, James Y. Zou, Jignesh Patel, Ronald Witteles, Susan Cheng, and David Ouyang. High-throughput precision phenotyping of left ventricular hypertrophy with cardiovascular deep learning. *JAMA Cardiology*, 7(4): 386–395, 2022. . URL https://doi.org/10.1001/jamacardio.2021.6059.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 203–213. IEEE/CVF, 2020. .

Zhiyuan Gao, Dominic Yurk, and Yaser S. Abu-Mostafa. Machine learning with scarce data: Ejection fraction prediction using plax view. In *Proceedings of the Medical Imaging with Deep Learning Conference (MIDL)*, 2025a. URL https://openreview.net/forum?id=JEN5FzeFZj. Oral presentation. OpenReview submission ID: JEN5FzeFZj.

Zhiyuan Gao, Dominic Yurk, and Yaser S. Abu-Mostafa. Plax ef labels dataset, 2025b. Available at: https://github.com/Jeffrey4899/PLAX_EF_Labels_202509.

Jamie Alexis D. Goco, Mohammad H. Jafari, Christina Luong, Teresa Tsang, and Purang Abolmaesumi. An efficient deep landmark detection network for PLAX EF estimation using sparse annotations. In Cristian A. Linte and Jeffrey H. Siewerdsen, editors, *Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 12034, page 120340N. International Society for Optics and Photonics, SPIE, 2022. . URL https://doi.org/10.1117/12.2611239.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000. . URL https://www.ahajournals.org/doi/abs/10.1161/01.CIR.101.23.e215.

Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Alistair Johnson, Elizabeth Herbst, Jonathan W. Waks, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECHO: Echocardiogram Matched Subset (version 0.1), 2023. URL https://doi.org/10.13026/ef48-v217. RRID:SCR_007345; Published July 21, 2023.

Zhe Huang, Gary Long, Benjamin Wessler, and Michael C. Hughes. Tmed 2: A dataset for semi-supervised classification of echocardiograms. In *DataPerf workshop at ICML*, 2022. URL https://tmed.cs.tufts.edu/papers/HuangEtAl_TMED2_DataPerf_2022.pdf.

Alistair E. W. Johnson, Tom J. Pollard, Steven Horng, Leo Anthony Celi, and Roger G. Mark. Mimic-iv-note: Deidentified free-text clinical notes, 2023. URL https://physionet.org/content/mimic-iv-note/2.2/. Deidentified free-text clinical notes from MIMIC-IV clinical database.

Youngjun Kim, Jennifer H. Garvin, Mary K. Goldstein, Tammy S. Hwang, Andrew Redd, Dan Bolton, Paul A. Heidenreich, and Stéphane M. Meystre. Extraction of left ventricular ejection fraction information from various

types of clinical reports. *Journal of Biomedical Informatics*, 67:42–48, 2017. . URL https://doi.org/10.1016/j.jbi.2017.01.017.

Roberto M. Lang, Luigi P. Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A. Flachskampf, Elyse Foster, Steven A. Goldstein, Tatiana Kuznetsova, Patrizio Lancellotti, Denisa Muraru, Michael H. Picard, Ernst R. Rietzschel, Lawrence Rudski, Kirk T. Spencer, Teresa S. M. Tsang, and Jens-Uwe Voigt. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Journal of the American Society of Echocardiography*, 28(1):1–39.e14, 2015. . URL https://www.asecho.org/wp-content/uploads/2016/02/2015_ChamberQuantificationREV.pdf.

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Östvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography, 2019. URL https://arxiv.org/abs/1908.06948.

Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*, 2013. URL https://www.semanticscholar.org/paper/Pseudo-Label%3A-The-Simple-and-Efficient-Learning-Lee/798d9840d2439a0e5d47bcf5d164aa46d5e7dc26.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. . URL https://doi.org/10.1038/s41586-020-2145-8.

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://arxiv.org/abs/1605.07723.

Frances M. Russell, Audrey Herbert, David Manring, Matt A. Rutz, Benjamin Nti, Loren K. Rood, and Robert R. Ehrman. The parasternal long axis view in isolation: Is it good enough? *Journal of Emergency Medicine*, 62(6):769–774, 2022. . URL https://pubmed.ncbi.nlm.nih.gov/35562250/.

Erik Smistad, Andreas Ostvik, Ivar Mjaland Salte, Daniela Melichova, Thuy Mi Nguyen, Kristina Haugaa, Harald Brunvand, Thor Edvardsen, Sarah Leclerc, Olivier Bernard, Bjørnar Grenne, and Lasse Lovstakken. Real-time automatic ejection fraction and foreshortening detection using deep learning. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67(12):2595–2604, 2020. . URL https://pubmed.ncbi.nlm.nih.gov/32175861/.

Louis E. Teichholz, Thomas Kreulen, Mark V. Herman, and Richard Gorlin. Problems in echocardiographic volume determinations: Echocardiographic-angiographic correlations in the presence or absence of asynergy. *The American Journal of Cardiology*, 37(1):7–11, 1976. . URL https://pubmed.ncbi.nlm.nih.gov/1244736/.

Takeshi Tohyama, Ahram Han, Dukyong Yoon, Kenneth Paik, Brian Gow, Nura Izath, Jacques Kpodonu, and Leo Anthony Celi. Multi-view echocardiographic embedding for accessible ai development. *medRxiv*, 2025. . URL https://doi.org/10.1101/2025.08.15.25333725. Preprint. Epub 19 August 2025.

Connie W. Tsao, Aaron W. Aday, Zaid I. Almarzooq, Cheryl A.M. Anderson, Pankaj Arora, Christy L. Avery, Carissa M. Baker-Smith, Andrea Z. Beaton, Amelia K. Boehme, Alfred E. Buxton, Yvonne Commodore-Mensah, Mitchell S.V. Elkind, Kelly R. Evenson, Chete Eze-Nliam, Setri Fugar, Giuliano Generoso, Debra G. Heard, Swapnil Hiremath, Jennifer E. Ho, Rizwan Kalani, Dhruv S. Kazi, Darae Ko, Deborah A. Levine, Junxiu Liu, Jun Ma, Jared W. Magnani, Erin D. Michos, Michael E. Mussolino, Sankar D. Navaneethan, Nisha I. Parikh, Remy Poudel, Mary Rezk-Hanna, Gregory A. Roth, Nilay S. Shah, Marie-Pierre St-Onge, Evan L. Thacker, Salim S. Virani, Jenifer H. Voeks, Nae-Yuh Wang, Nathan D. Wong, Sally S. Wong, Kristine Yaffe, Seth S. Martin, on behalf of the American Heart Association Council on Epidemiology, Prevention Statistics Committee, and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2023 update: A report from the american heart association. *Circulation*, 147(8):e93–e621, 2023. . URL https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000001123.

Roberto Vega, Cherise Kwok, Abhilash Rakkunedeth Hareendranathan, Arun Nagdev, and Jacob L. Jaremko. Assessment of an artificial intelligence tool for estimating left ventricular ejection fraction in echocardiograms from apical and parasternal long-axis views. *Di-*

agnostics, 14(16), 2024. ISSN 2075-4418. . URL https://www.mdpi.com/2075-4418/14/16/1719.

Roberto Vega, Arun Nagdev, Masood Dehghan, Seyed Ehsan Seyed Bolouri, Brian Buchanan, Jeevesh Kapur, and Jacob L. Jaremko. A wall tracking method to estimate ejection fraction from the parasternal long-axis view in point of care ultrasound. *Ultrasound Open*, 3:100097, 2025. . URL https://doi.org/10.1016/j.wfumbo.2025.100097.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification, 2020. URL https://arxiv.org/abs/1911.04252.

Y. Zheng, M. Zhang, N. Yao, B. Huang, Y. Niu, and H. Chen. GL-Fusion: A global-local fusion network for multi-view echocardiogram segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 78–88. Springer, 2023. . URL https://doi.org/10.1007/978-3-031-43901-8_8.

Appendix A. Training Details

This appendix documents implementation details and hyperparameters for all trainable components in our pipeline. The goal is to enable faithful reproduction while keeping the main text focused on methodology and results.

A.1 Video View Classifier Training Details

This subsection summarizes training and implementation details for the view-classification components used to mine A4C and PLAX clips from MIMIC-IV-Echo (Section 4.1). We focus on model training and data preprocessing details; the manual reviewer verification results are reported in the main text.

A.1.1 TMED-2 Image View Classifier (ResNet-34)

Task and labels. We first train an image-based view classifier on TMED-2 (Huang et al., 2022) to bootstrap high-confidence view candidates from unlabeled MIMIC-IV-Echo videos. The classifier predicts five categories: A4C, A2C, PLAX, PSAX, and a combined "A4C/A2C/OTHER" class. In our implementation, string labels are mapped to class indices as follows: `A4C`$\rightarrow 0$, `A2C`$\rightarrow 1$, `PLAX`$\rightarrow 2$, `PSAX`$\rightarrow 3$, and `A4CorA2CorOther`$\rightarrow 4$.

Model and architectural adjustments. We use a ResNet-34 backbone initialized with ImageNet weights. To better match the relatively small echocardiography frame resolution and avoid overly aggressive early downsampling, we replace the first convolution with a 3×3 kernel (stride 1, padding 1) and remove the initial max-pooling layer. The final fully connected layer is replaced with a linear classifier outputting 5 logits.

Preprocessing. Each frame is resized to 112×112. Because TMED-2 images may be stored as single-channel intensity images, we convert inputs to 3 channels (grayscale replication) and apply standard ImageNet normalization after converting to a tensor.

Label smoothing and loss. To improve robustness to label noise and the heterogeneous "A4C/A2C/OTHER" category, we apply label smoothing with $\epsilon = 0.05$. Specifically, each hard label is converted into a 5-dimensional target distribution where the ground-truth class receives probability $1-\epsilon$ and the remaining mass ϵ is distributed uniformly over the other classes. The model output logits are passed through `log_softmax`, and we optimize the KL divergence between predicted log-probabilities and the smoothed target distribution using KLDivLoss with `reduction=batchmean`.

Optimization and training protocol. We follow the official TMED-2 split indicators provided in our per-image CSV metadata, training on the `train` subset and monitoring loss on the `val` subset. Optimization uses SGD with momentum 0.9, initial learning rate 10^{-3}, batch size 32, and 50 epochs. We apply a StepLR schedule with step size 10 and decay factor 0.1. The checkpoint with the lowest validation loss is saved as the final image model.

Implementation notes and compute. Training is implemented in PyTorch with `DataParallel`. In our setup, the training runs on one NVIDIA A100 GPUs. Data loading uses 1 worker process. Validation curves and learning-rate schedules are logged via Weights & Biases for traceability.

A.1.2 Video View Classifier (X3D-s)

Training data. We train a three-way video classifier for `A4C/PLAX/OTHER` using the multi-dataset strategy described in Section 4.1: A4C videos from EchoNet-Dynamic (Ouyang et al., 2020), PLAX videos from EchoNet-LVH (Duffy et al., 2022), and `OTHER` videos mined from MIMIC-IV-Echo using the TMED-2 image classifier with conservative score thresholds to reduce contamination.

Clip sampling and preprocessing. Videos are decoded into RGB frames and converted into fixed-length clips with $L=64$ frames using temporal stride 2. For sufficiently long videos, a random start frame is sampled to provide temporal augmentation; if the decoded clip is shorter than L, missing frames are padded with black frames to ensure a fixed temporal input length. Frame-level transforms (e.g., padding/cropping and normalization) are applied independently to each frame and then stacked into a $L\times C\times H\times W$ tensor before being permuted to the $C\times L\times H\times W$ format expected by 3D backbones.

Model and training. We fine-tune a pretrained X3D-s network (Feichtenhofer, 2020) by replacing the final projection layer with a three-logit classifier head for `A4C/PLAX/OTHER`. The training objective is standard multi-class classification, and we select the final checkpoint based on validation performance on a held-out split.

Implementation notes and compute. Video-model training is implemented in PyTorch and run on three NVIDIA A100 GPUs using data-parallel training. Data loading uses 8 worker processes. Training progress and validation metrics are logged via Weights & Biases.

A.2 A4C EF Regressor (EchoNet-Dynamic)

Backbone and regression head. We use the TorchVision `r2plus1d_18` R(2+1)D architecture initialized from a public video-pretrained checkpoint. The final classification layer is replaced with a single-neuron linear layer to output a scalar EF prediction.

Clip sampling and preprocessing. Given an input video, we decode RGB frames and construct a fixed-length clip of

L=64 frames using temporal stride 2 (i.e., sampling every other frame from a 128-frame window). For sufficiently long videos, the start frame is randomly sampled to provide temporal augmentation; for shorter clips, missing frames are padded with black frames to maintain a consistent input length. Each frame is converted to channel-first format and normalized after scaling to $[0, 1]$. Spatial augmentation follows a lightweight padding-and-cropping strategy at a target resolution of 112×112: with equal probability, we either (i) pad by 24 pixels and crop a fixed top-right 112×112 region, or (ii) pad by 12 pixels and apply a random 112×112 crop. During evaluation, we disable spatial augmentation and use only normalization.

Optimization and model selection. We train on the official EchoNet-Dynamic `TRAIN` split, monitor performance on `VAL`, and report final performance on `TEST` (75%/12.5%/12.5%). The training objective is mean squared error (MSE) between predicted and ground-truth EF values (percentage points). We optimize with RAdam (initial learning rate 10^{-3}, batch size 32) for 100 epochs. A ReduceLROnPlateau scheduler (patience 5 epochs, reduction factor 0.1) adapts the learning rate based on validation loss. The checkpoint with the lowest validation loss is selected as the final A4C model.

Implementation notes. All models are implemented in PyTorch and trained on a single NVIDIA H100 GPU. Data loading uses 8 worker processes. Training progress, including validation metrics and learning-rate schedules, is logged using Weights & Biases to ensure traceability and experiment reproducibility.

A.3 PLAX EF Model Training Details

This subsection documents implementation details for the PLAX EF regression models reported in Section 5.1. We train two models with identical architectures and training protocols, differing only in batch size (16 vs. 32); the final PLAX predictor is formed by simple unweighted ensembling.

Training data and proxy targets. PLAX training clips are constructed as described in Section 4.5. Each PLAX video inherits a *study-level* proxy EF label computed by averaging the A4C teacher model predictions across all A4C clips available in the same study. The resulting PLAX training set is split into `TRAIN`/`VAL` at the study level (Section 4.5) so that multiple clips from the same study do not leak across splits.

Architecture. Both PLAX models use the TorchVision `r2plus1d_18` R(2+1)D backbone initialized from a public video-pretrained checkpoint. We replace the final classification layer with a single-neuron linear head to output a scalar EF prediction.

Clip sampling and preprocessing. Given an input PLAX video, we decode RGB frames and construct a fixed-length clip of L=64 frames using temporal stride 2. For sufficiently long videos, the start frame is randomly sampled to provide temporal augmentation; for shorter videos, missing frames are padded with black frames to maintain a consistent temporal input length. Frame-wise preprocessing is applied before stacking frames into a clip tensor. In our implementation, we use padding followed by random cropping to a fixed spatial resolution, together with intensity scaling to $[0, 1]$ and channel-wise normalization.

Optimization and model selection. We train both models using mean squared error (MSE) loss between predicted EF and proxy EF labels. Optimization uses the RAdam optimizer with initial learning rate 10^{-3} and a maximum of 100 epochs. We apply a ReduceLROnPlateau learning rate scheduler (patience 5 epochs, reduction factor 0.1) based on validation loss. The checkpoint with the lowest `VAL` loss is selected for each configuration. The two configurations reported in Table 2 differ only by batch size (16 vs. 32); all other settings are kept fixed.

Study-level inference and ensembling. At inference time, we compute study-level predictions by averaging video-level EF predictions across all PLAX clips within each study. The final PLAX estimate is then obtained by an unweighted average of the two study-level predictions from the batch-size-16 and batch-size-32 models (i.e., a two-model ensemble).

Implementation notes and compute. Training is implemented in PyTorch. The PLAX models are trained on a single NVIDIA H100 GPU, with data loading using 8 worker processes. Training curves and validation metrics are logged via Weights & Biases for traceability.

Appendix B. Artifact Access and Usage

This appendix summarizes what is released and how to use the artifacts listed in Section 6. The released resources primarily support *inference and evaluation* with pretrained models and released label files. We do not redistribute raw MIMIC-IV-Echo videos due to the PhysioNet data-use agreements. An overview of the released artifacts is provided in Table 4, and a screenshot of the Hugging Face Space demo interface is shown in Fig. 12.

B.1 Overview of released artifacts

Table 4 provides a concise overview of the released artifacts. URLs are listed in Section 6 to avoid duplication.

Table 4: Overview of released artifacts (URLs are listed in Section 6).

Artifact	Summary
GitHub repository	CSV label files for reconstructing PLAX/A4C cohorts in MIMIC-IV-Echo (proxy-labeled PLAX train/val, note-derived ground-truth cohorts, and the paired multi-view subset), plus EF-extraction prompts/notebooks and instructions for locating videos under authorized access.
Hugging Face model weights	Pretrained checkpoints for A4C EF inference and PLAX EF inference. The main paper uses a two-checkpoint PLAX ensemble (batch sizes 16 and 32) and averages their study-level predictions.
Hugging Face Space demo	Browser-based UI for uploading short PLAX and/or A4C clips and obtaining per-view EF predictions, as well as their mean when both views are provided (research and education use only).
Google Colab notebook	Runnable notebook that downloads released weights and demonstrates inference on sample clips or user-provided uploads; can be adapted to compute metrics given authorized access to the underlying datasets.

B.2 Hugging Face Space demo interface

Fig. 12 shows the Hugging Face Space interface used for online inference. Users may upload a PLAX clip and/or an A4C clip and obtain `EF from A4C`, `EF from PLAX`, and `EF (mean of available views)`. The Space is intended for research and education use only and is not a medical device.

B.3 Reproducing the reported evaluation

To reproduce the metrics reported in Section 5, the required steps are:

- **Obtain authorized access.** Obtain credentialed access to MIMIC-IV-Echo and (if needed) MIMIC-IV-Note under the PhysioNet data-use agreements.
- **Reconstruct cohorts from label files.** Use the released CSV label files to locate the corresponding videos in MIMIC-IV-Echo (e.g., via `subject_group`, `subject_id`, `study_id`, and `file_id`, depending on the label format).
- **Run inference with released weights.** Download the pretrained A4C and PLAX model checkpoints and run inference on the target clips.
- **Compute study-level metrics.** Aggregate video-level predictions by averaging within each study (as in Section 5) and report MAE and correlation on the note-derived ground-truth cohort.

B.4 Running inference without local setup

Hugging Face Space. Open the Space URL (Section 6), upload a PLAX clip and/or an A4C clip (short MP4 recommended), and click `Submit`. The demo reports `EF from A4C`, `EF from PLAX`, and the mean when both are available.

Google Colab. Open the Colab notebook URL (Section 6) and run all cells. The notebook downloads the released weights and runs inference on sample inputs or user-provided clips. For simplicity, lightweight demos may use one PLAX model and one A4C model, whereas the main paper results use the full aggregation setup (two PLAX checkpoints plus the A4C model, as described in the paper).

B.5 Practical notes

- **Input clip length and format.** The Hugging Face demo is intended for short clips (e.g., ≤ 10s) in MP4 format. Very long clips may time out or increase latency in online inference environments.
- **View requirements.** The released models are trained for A4C and PLAX views. If the uploaded clip is not a true A4C/PLAX view (or is a partial/atypical acquisition), predictions may be unreliable. For rigorous evaluation, we recommend using view-filtered cohorts and study-level aggregation as described in Section 5.
- **Meaning of demo outputs.** `EF from A4C` and `EF from PLAX` are per-view EF predictions produced by the corresponding view-specific models. `EF (mean of available views)` is a simple unweighted average of the available view predictions (if only one view is provided, it equals that view's estimate).
- **Study-level aggregation.** The main paper reports *study-level* results by averaging predictions over multiple clips within the same study. Single-clip predictions (e.g., via the demo) are best interpreted as qualitative examples rather than a substitute for study-level reporting.
- **Privacy and intended use.** All artifacts are intended for research and education use only and must not be used for

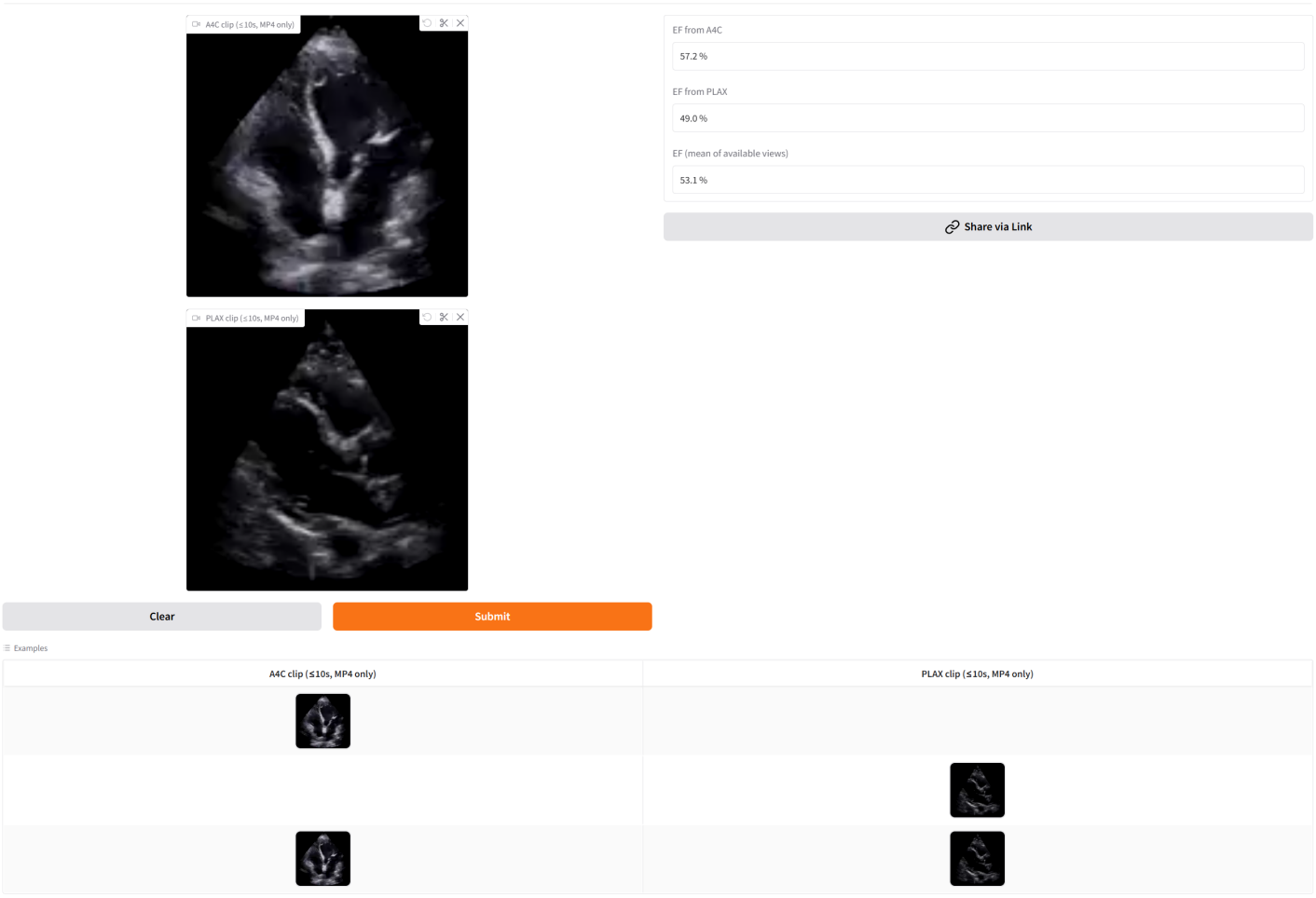


Figure 12: Screenshot of the Hugging Face Space demo for EF estimation from PLAX and A4C clips. The demo reports `EF from A4C`, `EF from PLAX`, and `EF (mean of available views)` when both are provided. The demo is intended for research and education use only and is not a medical device.

clinical diagnosis or treatment. Users are responsible for ensuring that any uploaded data (e.g., to online demos) comply with applicable privacy policies, regulations, and data-use agreements. When in doubt, prefer running inference locally or in Colab on appropriately de-identified data.

B.6 Notes on data redistribution and intended use

Raw MIMIC-IV-Echo videos are not redistributed due to data-use agreements. The released label files provide identifiers and instructions for locating the corresponding samples under authorized access. All released artifacts are intended for research and education use only and are not medical devices.