ECHOTAB: LEARNING MEASUREMENT-GROUNDED REPRESENTATIONS FOR ECHOCARDIOGRAPHIC DIAGNOSIS THROUGH TABULAR RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnostic information in multi-view medical imaging studies, like echocardiography, is distributed across views, each capturing different aspects of cardiac structure and function. In clinical practice, these views are first interpreted through quantitative measurements, which then support specific diagnoses. However, existing echo representation learning methods do not explicitly model measurements as an intermediate representation between views and diagnosis. As a result, individual measurements are not explicitly represented in a form that downstream diagnostic tasks can selectively use, hindering accurate diagnosis.

Inspired by this hierarchical clinical workflow, we introduce ECHOTAB, a measurement-grounded representation learning framework that represents multi-view echocardiography studies using measurement-specific tokens as an intermediate representation between echo views and diagnosis. Each measurement token selectively attends to the views relevant to its estimation and is trained through tabular reconstruction to predict the corresponding measurement value. The resulting tokens form a factorized echo representation, which encode clinically meaningful structural and functional properties and can be selectively used for downstream diagnosis. Unlike the commonly used contrastive approaches, ECHOTAB does not require negative samples or large batch sizes. Experiments on two datasets demonstrate that ECHOTAB learns effective measurement-grounded representations that improve diagnosis classification in multi-view echocardiography and its learned attention aligns with established measurement-view relationships used by cardiologists. Our code is accessible at `https://anonymous.4open.science/r/EchoTab`.

1 INTRODUCTION

Echocardiography (echo) is a common cardiac modality to assess the structure and function of the heart. Echocardiography interpretation follows a hierarchical process in which echo views in an echo study give rise to quantitative measurements that inform clinical diagnosis. Importantly, different measurements are derived from different subsets of views, and different diagnoses are informed by specific subsets of measurements. For example, the left ventricular ejection fraction measurement is derived from apical two- (A2C) and four-chamber (A4C) views and is then used to assess the severity of left ventricular systolic dysfunction.

Despite this hierarchical structure, existing echo foundation models are trained directly (Holste et al., 2025) or must be finetuned (Munim et al., 2026; Zhang et al., 2025) for diagnosis prediction, or leverage retrieval of text reports that contain diagnoses (Vukadinovic et al., 2026). However, categorical diagnosis labels introduce a discontinuity into the learning target and lose information about how strongly a patient expresses a given phenotype. Instead, measurements provide continuous, dense supervision. Rather than requiring the model to learn a thresholded clinical decision, measurement prediction encourages it to capture the anatomical and functional properties that underlie that decision. Some existing approaches already incorporate measurements but do not explicitly represent them as continuous targets. For example, EchoPrime (Vukadinovic et al., 2026) uses echo reports for contrastive learning, indirectly exposing the model to measurements described in the text. However, these measurements are encoded by the text encoder as discrete text representations rather than

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

as continuous values. A natural alternative is to use measurements directly as the target modality for contrastive learning. However, contrastive learning requires the set of measurements to be encoded into a single tabular embedding, which is then aligned with a single echo embedding. While this makes the resulting echo representations measurement-aware, the individual measurements are not explicitly represented and therefore cannot be factorized within the learned representations.

Motivated by the clinical hierarchy of echocardiographic interpretation, we introduce ECHOTAB, a measurement-grounded representation learning framework that models measurements as an intermediate representation between echo views and downstream diagnosis. Rather than collapsing measurements into a single embedding, ECHOTAB assigns a dedicated token to each measurement, each selectively attending to the relevant echo views. These measurement-specific tokens thus form a factorized representation of the echo study. A prediction head is trained to reconstruct the corresponding measurements from these tokens, encouraging each token to capture clinically meaningful structural and functional properties. These measurement-specific tokens can then be selectively combined for accurate downstream diagnosis. Unlike contrastive approaches, ECHOTAB uses a tabular reconstruction objective, which does not require negative samples and large batch sizes.

Our contributions are:

- We propose ECHOTAB, a measurement-grounded representation learning framework that represents multi-view echocardiography studies using measurement-specific tokens, which serve as an intermediate representation between echo views and diagnosis.
- We show that ECHOTAB learns these measurement-specific tokens through tabular reconstruction. This enables diagnosis-specific selection of relevant measurements, improving diagnosis performance on two datasets over pre-training methods that learn generic study-level representations, while avoiding the need for negative samples and large batch sizes.
- We systematically study ECHOTAB's architecture and behavior, including the impact of different measurement prediction heads, and show that its learned attention over echo views is consistent with established measurement-view relationships used by cardiologists.

2 ECHOTAB

We propose ECHOTAB, a measurement-grounded representation learning framework for multi-view echo studies. As shown in Figure 1, given an echo study, each individual view is first encoded by the echo encoder (Section 2.1), resulting in a set of echo tokens per view. A tabular decoder with measurement queries (Section 2.2) then associates each measurement with the echo views most relevant to it, producing measurement tokens that the prediction head maps to the final measurements (Section 2.3). The framework is trained end-to-end with a tabular reconstruction objective (Section 2.4), after which relevant measurement tokens can be selected for accurate downstream diagnosis.

2.1 ECHO ENCODER

An echocardiography study $S = \{v_1, \ldots, v_N\}$ consists of N video clips, each assigned a view label $\kappa(v_i) \in \mathcal{C}$, where $\mathcal{C}$ denotes the set of view types (e.g., A4C and PLAX). Let $\mathcal{V} \subseteq \mathcal{C}$ denote the subset of view types present in a given study. For each $c \in \mathcal{V}$, we select exactly one video clip $v^{(c)} \in S$ satisfying $\kappa(v^{(c)}) = c$. During training, the selected clip for each available view is sampled uniformly at random from all clips in S with label c. At inference, the first such clip is selected deterministically.

The echo encoder is initialized with EchoPrime (Vukadinovic et al., 2026) weights and uses the MViTv2 (Li et al., 2022) architecture. Each selected view instance $v^{(c)}$ is encoded independently, and we retain $L_{v^{(c)}}$ local patch tokens $\{z_{v^{(c)},j}\}_{j=1}^{L_{v^{(c)}}}$ rather than the global `CLS` token, where each $z_{v^{(c)},j}$ represents a spatio-temporal patch of the video clip. The local tokens from all selected views are concatenated into the study-level echo token sequence,

$$Z = \left[z_{v^{(c)},j} + e_c, \right]_{c \in \mathcal{V}, j=1,\ldots,L_{v^{(c)}}}, \tag{1}$$

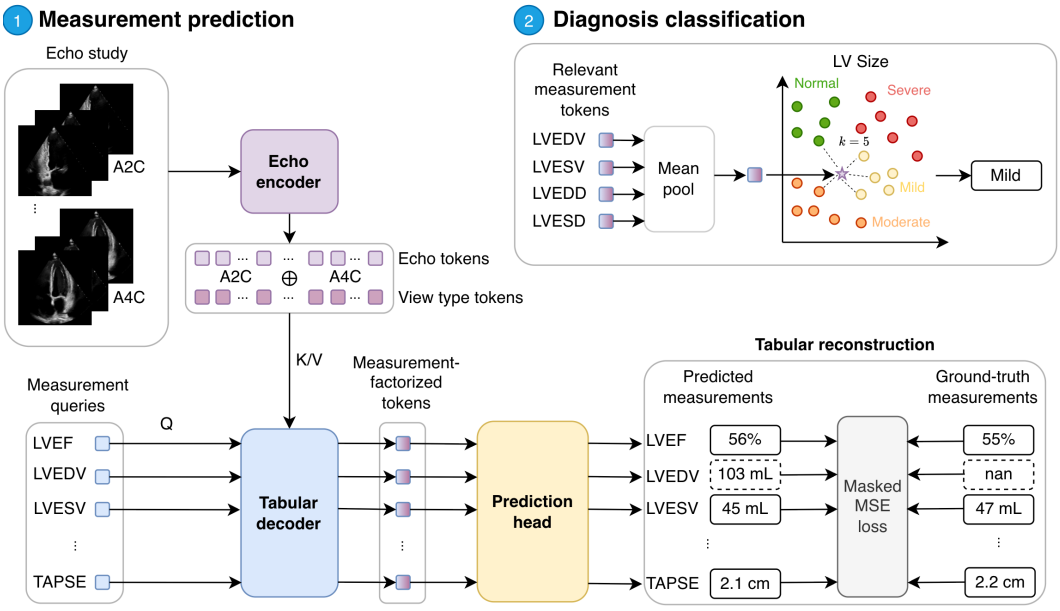


Figure 1: **Overview of ECHOTAB, our measurement-grounded representation learning framework for multi-view echo studies.** (1) Given an echo study, each individual view is first encoded by the echo encoder. Next, a tabular decoder lets each measurement query selectively pull information from the echo views that are relevant for estimating that measurement. This results in a set of measurement-factorized tokens for the echo study, which the prediction head uses to predict measurement values. ECHOTAB is trained with a tabular reconstruction objective, in which the MSE loss between available ground-truth and predicted measurements is calculated. (2) ECHOTAB's measurement-factorized tokens enables accurate diagnosis by using only the measurement tokens that are relevant for the downstream diagnosis.

where $e_c \in \mathbb{R}^d$ is a learnable view embedding for view type c. In the subsequent step, we mask background patch tokens, which are those with mean pixel intensity below the threshold $\tau = -0.6$ (corresponding to approx. 0 before normalization). Although MViTv2's (Li et al., 2022) multiscale architecture already endows each local token with spatial context beyond its own patch, we empirically find that masking background tokens improves view classification performance (App. D.3), suggesting that representations based primarily on foreground regions are more discriminative for view identification.

2.2 TABULAR DECODER

Each measurement in an echo study is derived from a subset of echo views. For example, the left ventricular ejection fraction measurement is estimated from the apical two-chamber (A2C) and four-chamber (A4C) views. We therefore design the tabular decoder to allow each measurement token to selectively aggregate information from the echo views relevant to its estimation.

The tabular decoder consists of B cross-attention blocks, each followed by a feed-forward network (FFN), with residual connections, layer normalization, and learnable layer-scale parameters. We initialize learnable measurement queries $q_m \in \mathbb{R}^d$ for measurement $m \in \{1, \ldots, M\}$. These serve as queries, while the echo token sequence from Section 2.1 serve as keys and values. Before entering the decoder, the two token sets are mapped to a common dimension d using separate linear projections. The tabular decoder transforms each q_m into a measurement-factorized representation $x_m \in \mathbb{R}^d$, which is subsequently used to predict measurement m. We omit self-attention between measurement tokens to keep the resulting measurement representations independent of one another, yielding a factorized representation for downstream diagnosis.

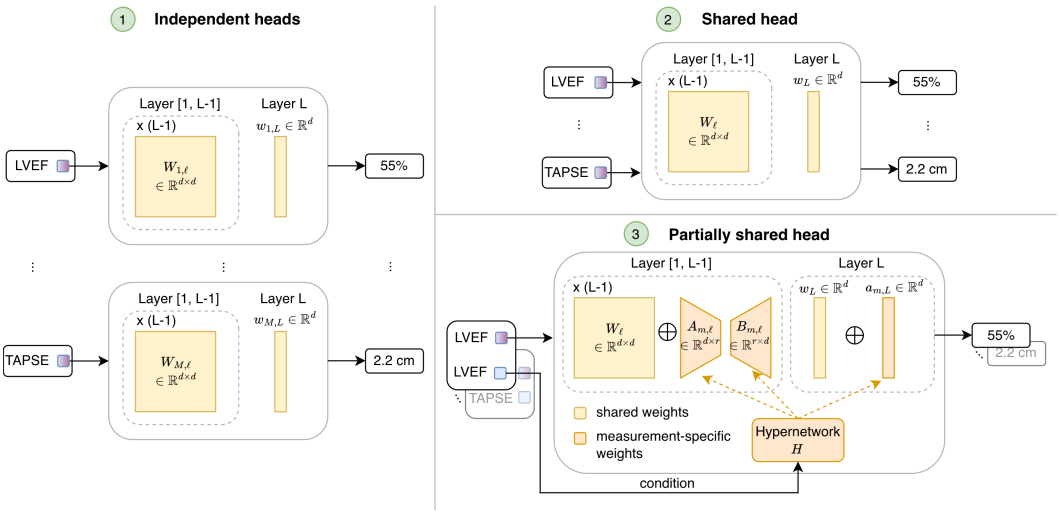


Figure 2: **Variants of the prediction head used in ECHOTAB**. The prediction head is used during the tabular reconstruction of continuous measurements from their corresponding measurement-specific representations. We compare different degrees of parameter sharing to assess the trade-off between shared and measurement-specific components of the prediction function. (1) **Independent heads**: each measurement token has its own MLP head, enabling fully measurement-specific parameters. (2) **Shared head**: all measurement tokens share the same MLP head, enforcing fully shared parameters. (3) **Partially shared head**: all measurement tokens share the base weights of the MLP head, while a hypernetwork generates measurement-specific parameter updates, providing a compromise between shared and measurement-specific parameters. Biases are omitted for clarity.

2.3 PREDICTION HEAD

The prediction head acts as a lightweight read-out on top of the tabular decoder in Section 2.2: given the measurement-factorized tokens from the tabular decoder, it reconstructs the corresponding tabular measurement values. As shown in Figure 2, we propose three variants of the prediction head: independent, shared, and partially shared. Throughout, $\sigma(\cdot)$ denotes the chosen element-wise non-linearity (e.g., GELU), and $h_m^{(0)} = x_m$.

Independent heads. This is the standard multi-task setting, where each measurement token is assigned a separate L-layer MLP head. For measurement $m \in \{1, \dots, M\}$,

$$\begin{aligned} h_m^{(\ell)} &= \sigma\left(W_{m,\ell} h_m^{(\ell-1)} + b_{m,\ell}\right), \quad \ell = 1, \dots, L-1, \\ \hat{y}_m &= w_{m,L}^\top h_m^{(L-1)} + b_{m,L}, \end{aligned} \tag{2}$$

with $W_{m,\ell} \in \mathbb{R}^{d \times d}$, $b_{m,\ell} \in \mathbb{R}^d$, $w_{m,L} \in \mathbb{R}^d$, and $b_{m,L} \in \mathbb{R}$. Thus, each measurement learns an independent set of prediction-head parameters, and the total number of parameters scales linearly with the number of measurements, i.e., $O(M)$.

Shared head. A single L-layer MLP head is shared across all measurement tokens. For measurement $m \in \{1, \dots, M\}$,

$$\begin{aligned} h_m^{(\ell)} &= \sigma\left(W_\ell h_m^{(\ell-1)} + b_\ell\right), \quad \ell = 1, \dots, L-1, \\ \hat{y}_m &= w_L^\top h_m^{(L-1)} + b_L, \end{aligned} \tag{3}$$

with $W_\ell \in \mathbb{R}^{d \times d}$, $b_\ell \in \mathbb{R}^d$, $w_L \in \mathbb{R}^d$, and $b_L \in \mathbb{R}$. The same parameters are therefore used to predict every measurement, making the number of prediction-head parameters independent of the number of measurements, i.e., $O(1)$ w.r.t. M.

Partially shared head. Inspired by text-to-LoRA (Charakorn et al., 2025), which uses a hypernetwork (Ha et al., 2017) conditioned on a task description to generate task-specific low-rank corrections for an LLM, the partially shared head analogously treats the shared L-layer MLP as the base model and a hypernetwork conditioned on the measurement query q_m generates measurement-specific low-rank corrections on top of the shared weights. For measurement $m \in \{1, \ldots, M\}$,

$$\begin{aligned} h_m^{(\ell)} &= \sigma\Big(\big(W_\ell + \tfrac{\alpha}{r} A_{m,\ell} B_{m,\ell}\big)\, h_m^{(\ell-1)} + b_\ell + \beta_{m,\ell}\Big), \quad \ell = 1, \ldots, L-1, \\ \hat{y}_m &= \big(w_L + a_{m,L}\big)^\top h_m^{(L-1)} + b_L + \beta'_{m,L}, \end{aligned} \tag{4}$$

where $W_\ell \in \mathbb{R}^{d\times d}$, $b_\ell \in \mathbb{R}^d$, $w_L \in \mathbb{R}^d$, and $b_L \in \mathbb{R}$ are shared, while $A_{m,\ell} \in \mathbb{R}^{d\times r}$, $B_{m,\ell} \in \mathbb{R}^{r\times d}$ with $r \ll d$, $\beta_{m,\ell} \in \mathbb{R}^d$, $a_{m,L} \in \mathbb{R}^d$, and $\beta'_{m,L} \in \mathbb{R}$ are measurement-specific corrections generated by a hypernetwork H conditioned on q_m. Following (Hu et al., 2022), the low-rank update is scaled by $\frac{\alpha}{r}$, where r is the rank and α is a scaling hyperparameter. We investigate the impact of various hypernetwork variants in Table 13.

2.4 Tabular Reconstruction Objective

Echo measurements are continuous and sparse: each study reports only a subset of them. They also vary widely in unit and scale. We therefore train EchoTab with a masked, variance-normalized mean squared error. For a mini-batch in which $\mathcal{O}$ is the set of observed (study, measurement) pairs,

$$\mathcal{L} = \frac{1}{|\mathcal{O}|} \sum_{(i,m)\in\mathcal{O}} \frac{(\hat{y}_{i,m} - y_{i,m})^2}{s_m^2}, \tag{5}$$

where $\hat{y}_{i,m}$ and $y_{i,m}$ are the predicted and true values of measurement m in study i, and s_m^2 is the variance of measurement m over the training set. Dividing by s_m^2 puts all measurements on a comparable footing, so the loss is not dominated by measurements with large absolute values.

3 Experiments

3.1 Data

We train EchoTab on two datasets separately: MIMIC-IV-Echo v1.0 (Gow et al., 2026; Pollard et al., 2026) and an internal dataset. After data processing, MIMIC-IV-Echo contains 4,514 patients with 7,144 studies and our internal cohort 7,404 patients with 9,431 studies (App. A & E).

3.2 Baselines and Evaluation Strategy

We evaluate EchoTab on diagnosis classification tasks on two datasets, covering conditions ranging from valvular disease to systolic dysfunction (App. C). For each task, we train a k-nearest neighbors classifier with $k = 5$ on the resulting study-level embeddings. We compare EchoTab against existing echo foundation models and three classes of baselines trained on the tabular measurements available in each dataset. First, a contrastive baseline aligns paired echo and measurement representations, encoded by the echo encoder and the pre-trained TARTE tabular encoder (Kim et al., 2025) respectively, in a shared embedding space, following the image-tabular formulation of MMCL (Hager et al., 2023). Second, a supervised contrastive (SupCon) baseline uses tabular similarity to define positive and negative echo pairs, following the idea of TGV (Hasny et al., 2026) and inspired by the original SupCon (Khosla et al., 2020). Third, an MLP baseline adds independent measurement-prediction heads to the echo encoder and is trained using the tabular reconstruction objective of EchoTab. Details on these baselines are provided in App. H.

By default, all models treat an echo study as the sample unit. However, for the contrastive and SupCon baselines, we additionally include view-level training to maximize batch size. Both these baselines are trained using 8 A40 GPUs, whereas the MLP baseline and EchoTab use a single A100 GPU with 32 GB of memory. Hyperparameters used by EchoTab are detailed in App. F.

Table 1: **ECHOTAB's measurement-specific representations enable accurate diagnosis for multi-view echocardiography.** By enabling diagnosis-specific selection of relevant measurement tokens, ECHOTAB consistently outperforms echo foundation models, as well as study- and view-level contrastive and MLP baselines, across downstream diagnosis classification tasks. Unlike contrastive approaches, ECHOTAB does not require large batch sizes, enabling efficient training. Bold and underlined values denote the **best** and second-best results.

(a) Diagnosis classification results on MIMIC (AUC).

Method	Echo encoder	Echo-tabular pre-training		Inference config.	MIMIC									
		Role of tabular data	Echo sample unit	Measurement tokens	Valves			Size				Sys. function		Avg.
					AS	AR	MR	LA	RA	LV	RV	LVEF	RVEF	
Echo foundation models (Reference)														
PanEcho	PanEcho	–	–	–	.597	**.644**	.609	.708	.717	**.877**	.646	.775	.748	.702
EchoJEPA-L	EchoJEPA-L	–	–	–	.525	.589	.557	.628	.730	.644	.660	.630	.721	.632
EchoPrime	EchoPrime	–	–	–	.526	.558	.618	.700	.724	.705	.638	.736	.725	.659
Baselines														
Contrastive	EchoPrime	Tabular alignment target	view	–	.552	.546	.615	.708	.782	.759	.694	.737	.753	.683
Contrastive	EchoPrime	Tabular alignment target	study	–	.583	.547	.618	.725	.782	.789	.737	.693	.732	.690
SupCon	EchoPrime	Tabular guidance	view	–	.672	.575	.599	.735	.778	.684	.704	**.844**	.749	.704
SupCon	EchoPrime	Tabular guidance	study	–	.592	.567	.599	.725	.754	.733	.743	.772	**.762**	.694
MLP	EchoPrime	Tabular reconstruction	study	–	.642	.551	.573	.713	.774	.745	.743	.775	.759	.697
Ours														
ECHOTAB	EchoPrime	Tabular reconstruction	study	Relevant	**.723**	.592	**.658**	**.745**	**.808**	.852	**.756**	.843	.754	**.748**

(b) Diagnosis classification results on the internal dataset (AUC).

Method	Echo encoder	Echo-tabular pre-training		Inference config.	Internal								
		Role of tabular data	Echo sample unit	Measurement tokens	Valves		Size				Sys. function		Avg.
					AS	MR	LA	RA	LV	RV	LVEF	RVEF	
Echo foundation models (Reference)													
PanEcho	PanEcho	–	–	–	.681	.685	.698	.750	.792	.672	.741	.593	.702
EchoJEPA-L	EchoJEPA-L	–	–	–	.606	.641	.682	.735	.600	.565	.666	.656	.644
EchoPrime	EchoPrime	–	–	–	.740	.711	.721	**.790**	.772	.660	.765	.684	.730
Baselines													
Contrastive	EchoPrime	Tabular alignment target	view	–	.734	.684	.760	.777	.791	.747	.762	.727	.748
Contrastive	EchoPrime	Tabular alignment target	study	–	.785	.696	.771	.777	.807	.697	.756	.692	.748
SupCon	EchoPrime	Tabular guidance	view	–	.777	.669	.779	.786	.816	.739	.801	.692	.757
SupCon	EchoPrime	Tabular guidance	study	–	**.809**	.696	**.792**	.760	**.830**	.703	.785	.627	.750
MLP	EchoPrime	Tabular reconstruction	study	–	.790	**.717**	.775	.765	.815	.685	.810	.731	.761
Ours													
ECHOTAB	EchoPrime	Tabular reconstruction	study	Relevant	.800	.713	.787	.766	**.830**	**.749**	**.829**	**.749**	**.778**

ECHOTAB factorizes the input echo study into measurement tokens $\{x_m\}_{m=1}^{M}$. For diagnosis k, let $\mathcal{M}_k$ denote the measurements relevant to its prediction (see Table 10 and Table 11). We construct a diagnosis-specific study embedding as

$$s_k = \frac{1}{|\mathcal{M}_k|} \sum_{m \in \mathcal{M}_k} x_m. \tag{6}$$

As the baselines do not provide this factorization into measurement-specific tokens, we instead use the study-level echo embedding obtained by averaging representations across available views.

In App. B.1, we investigate whether the advantage of ECHOTAB is solely from its diagnosis-specific prior, by giving the baselines an analogous prior in the form of diagnosis-relevant views. In App. B.2, we examine whether ECHOTAB benefits from restricting its input to these relevant views.

3.3 DIAGNOSIS CLASSIFICATION

Table 1 presents diagnosis classification results on MIMIC and our internal dataset. ECHOTAB consistently improves diagnosis classification across both datasets and achieves the highest average AUC among echo foundation models, contrastive baselines, and the MLP baseline. It ranks first or second on 8 of 9 diagnoses on MIMIC, and on 7 of 8 diagnoses on our internal dataset. Relative to EchoPrime (the shared initialization used by all baselines), ECHOTAB improves AUC on all 9

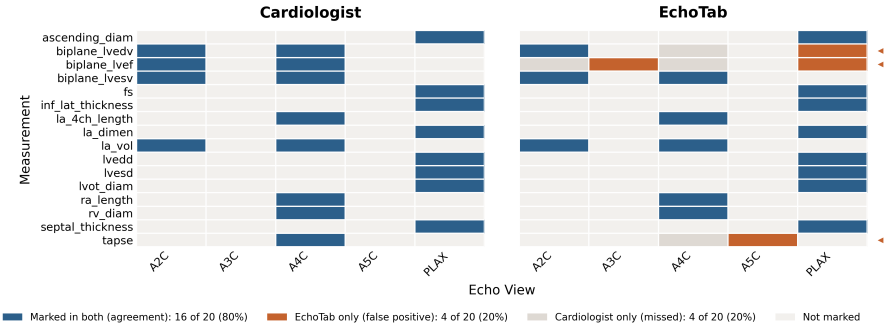


Figure 3: **Comparison of the echo views a cardiologist (left) and ECHOTAB (right) rely on when deriving measurements on MIMIC.** ECHOTAB matches the cardiologist's view in 16 of 20 (80%) measurement-view pairs, indicating it learns clinically meaningful measurement-view associations.

MIMIC diagnoses, increasing the average AUC from 0.659 to 0.748 (+13.5%). On the internal dataset, it improves performance on 7 of 8 diagnoses and raises the average AUC from 0.730 to 0.778 (+6.6%). All baselines also improve upon EchoPrime in average AUC, indicating that echo-tabular pre-training improves diagnosis accuracy.

On MIMIC, ECHOTAB outperforms the strongest baseline (view-level SupCon) by 4.4 average AUC points (0.748 vs. 0.704), while requiring only 1 GPU compared with 8 GPUs for SupCon. On the internal dataset, it outperforms the strongest baseline (MLP) on 7 of 8 diagnoses and improves average AUC by 1.7 points (0.778 vs. 0.761), despite using the same tabular reconstruction objective. This suggests that ECHOTAB's gains are not attributable only to the training objective, but also to its usage of diagnosis-relevant measurement tokens for accurate prediction (Section 4.1 for more).

3.4 ATTENTION MAP ANALYSIS

ECHOTAB's tabular decoder allows us to examine which echo views it attends to for individual measurements through its cross-attention maps. Figure 3 compares its attention pattern with cardiologist-identified relevant views on MIMIC. We first average the attention maps across all cross-attention layers and then binarize the resulting map to match the number of views deemed relevant by the cardiologist: if x views are considered relevant, the x highest-attended views are labeled relevant. Measurements that depend on view types absent from the view set $\mathcal{C}$ or on Doppler views are excluded. The latter are omitted because the view classifier groups all Apical Doppler acquisitions into a single broad category, which is too coarse for this measurement-view analysis.

ECHOTAB selects the same echo views as the cardiologist in 16 of 20 (80%) measurement-view pairs, agreeing on 13 of 16 (81%) measurements. The three disagreements involve biplane LVEDV and LVEF (left ventricle measurements) and TAPSE (right ventricular function). For TAPSE, it favors A5C over A4C. For LVEDV and LVEF, it partly shifts attention from A2C/A4C to PLAX/A3C. In each case, the alternative views ECHOTAB attends to also capture the relevant cardiac structure. Because echo views contain partially redundant information, the model may exploit correlated signals rather than strictly follow clinical protocols. Overall, ECHOTAB's attention captures clinically meaningful measurement-view associations, with deviations remaining anatomically plausible.

4 ABLATIONS

4.1 ARCHITECTURAL ABLATION

In Table 2, we ablate the individual components of ECHOTAB. Specifically, we investigate the effect of the tabular decoder, compare measurement queries with a single query shared across all measurements, and evaluate the prediction head variants described in Section 2.3. When present, the tabular decoder uses two cross-attention layers, and all prediction heads are two-layer MLPs. For experiments using measurement queries, we use diagnosis-relevant measurement tokens for diagnosis classification.

With independent heads, adding the tabular decoder improves diagnosis classification over the MLP baseline on both MIMIC (0.697 vs. 0.720) and the internal cohort (0.761 vs. 0.766). This suggests that querying echo views through the tabular decoder produces a study-level representation that is more informative for diagnosis classification than simply averaging the echo-view embeddings.

When prediction heads are independent, measurement queries slightly degrade performance relative to a single query (0.709 vs. 0.720 on MIMIC; 0.764 vs. 0.766 on internal). We attribute this to a trade-off between measurement-specific representational flexibility and cross-measurement representation alignment: separate queries allow measurements to be represented independently, but independent heads do not enforce a shared representation space among the resulting tokens. Hence, averaging these tokens can produce a less coherent study embedding for diagnosis classification.

A shared prediction head addresses this issue by mapping all measurement tokens into a common value space, while still preserving their distinct query identities. This retains measurement-specific representations while restoring cross-measurement representation alignment, making their average more coherent. Diagnosis classification consequently improves (0.732 on MIMIC; 0.773 on internal). Without measurement-specific queries, a shared head yields degenerate results because the same representation is used for every measurement, preventing measurement-specific predictions.

The partially shared head provides a compromise between measurement-specific representational flexibility and cross-measurement alignment, yielding the best diagnosis classification in both cohorts (0.748 on MIMIC; 0.778 on internal), corresponding to relative gains of 7.3% and 2.2% over the MLP baseline. It also achieves the best measurement prediction on MIMIC (0.418) and is essentially tied on the internal cohort (0.445) with the fully shared head (0.447) and the MLP (0.446).

Method	Echo encoder	Tabular decoder	Tabular decoder queries	Prediction head	MIMIC		Internal	
					Diag. clf.	Meas. pred.	Diag. clf.	Meas. pred.
MLP	✓	✗	–	Independent	.697	.392	.761	<u>.446</u>
				Shared	.683	-16.944	.648	-14.799
	✓	✓	Single	Independent	.720	.380	.766	.422
				Shared	.628	-16.987	.624	-14.902
				Independent	.709	.401	.764	.415
ECHOTAB	✓	✓	Measurement	Shared	<u>.732</u>	<u>.412</u>	<u>.773</u>	**.447**
				Partially shared	**.748**	**.418**	**.778**	.445

Table 2: **Ablation of tabular decoder, queries, and prediction head in ECHOTAB.** Bold and underlined values denote the **best** and <u>second-best</u> results. Diagnosis classification is measured by AUC averaged across diagnoses and measurement prediction by R^2 averaged across measurements. In ECHOTAB, a partially shared head makes the resulting measurement tokens most effective, balancing cross-measurement representation alignment with measurement-specific representational flexibility. As a result, it achieves the best diagnosis performance on both datasets.

4.2 MEASUREMENT TOKENS FOR DIAGNOSIS CLASSIFICATION

In Table 3, we ablate measurement-token configurations for diagnosis classification. In the *Relevant* setting, we average diagnosis-relevant measurement tokens. In *All*, we average all measurement tokens available during training. As an *Incorrect* control, we replace each diagnosis-specific subset with an equal-sized set of measurements sampled uniformly at random (without replacement) from other diagnoses, excluding that diagnosis's measurements. All settings use a partially shared head.

Averaging all measurement tokens mixes in diagnosis-irrelevant information and reduces classification performance. Likewise, using incorrect tokens also degrades performance. These results indicate that the learned measurement tokens encode measurement-specific information and that choosing diagnosis-relevant measurement tokens is crucial for accurate diagnosis.

Measurement tokens	MIMIC	Internal
Diagnosis-relevant	**.748**	**.778**
All	.693	.751
Incorrect	.687	.757

Table 3: **Ablation of the measurement tokens used for diagnosis classification** (avg. AUC). Choosing diagnosis-relevant measurement tokens is essential for accurate diagnosis.

5 Related Work

Echo foundation models. To our knowledge, no existing echo foundation model explicitly incorporates measurements as an intermediate step for diagnosis during training. EchoJEPA (Munim et al., 2026) and Echo-Vision-FM (Zhang et al., 2025) use self-supervised masked latent and masked pixel prediction objectives respectively, without measurements. EchoPrime (Vukadinovic et al., 2026) incorporates measurements only indirectly via video-text contrastive learning. In this setting, measurements are encoded as discrete text tokens, and diagnosis relies on retrieving semantically similar reports rather than reasoning from measurements. PanEcho (Holste et al., 2025) jointly predicts measurements and diagnoses, but optimizes the two tasks in parallel rather than using measurements to derive diagnosis. In contrast, EchoTab explicitly uses measurement representations as an intermediate step between echo views and diagnosis.

Vision-tabular learning. Building on CLIP's image-text contrastive learning (Radford et al., 2021), MMCL (Hager et al., 2023) aligns paired image and tabular embeddings in a shared latent space, encouraging the image encoder to capture information reflected in the associated tabular features. TGV (Hasny et al., 2026) instead uses tabular data to form positive and negative image pairs, encouraging the image encoder to learn representations that reflect tabular similarity. Both methods rely on contrastive objectives, which generally benefit from a larger and more diverse set of negative samples, obtained through larger batch sizes. In contrast, EchoTab uses a tabular reconstruction objective that requires no negative samples and therefore does not rely on large batches.

Concept bottlenecks and query-based decoders. Concept bottleneck models (Koh et al., 2020) and their variants (Espinosa Zarlenga et al., 2022; Oikarinen et al., 2023; Yuksekgonul et al., 2023) predict a set of semantically meaningful concepts from the input and then predict the target from those concepts, yielding an interpretable intermediate representation. Similar to concepts, the queries in query-based decoders extract information from an input, usually through attention layers. For example, object queries in DETR (Carion et al., 2020), output queries in Perceiver IO (Jaegle et al., 2022), and the learned queries of the Q-Former in BLIP-2 (Li et al., 2023). EchoTab combines these two ideas. Like a concept bottleneck, EchoTab uses measurement tokens as concepts that represent multi-view echos, which are subsequently used to predict diagnoses. Like a query-based decoder, it reads out these tokens from multi-view echos through cross-attention layers.

6 Discussion and Limitations

Discussion. Our experiments demonstrate the value of using measurement tokens as an intermediate representation between echo views and diagnosis. This formulation factorizes a multi-view echo study into measurement-specific tokens, allowing each diagnosis to use the measurement tokens most relevant to its prediction. Using these diagnosis-relevant measurement tokens improves diagnostic accuracy compared with generic study-level representations produced by alternative pretraining methods. Moreover, because these representations are grounded in measurement values, they provide a more interpretable basis for diagnosis than a direct mapping from echo views to diagnosis.

While our study focuses on echocardiography, our representation learning framework is, in principle, modality-agnostic and applicable to multi-view imaging studies with associated measurements. Future work could investigate its application to other modalities, such as cardiac MRI, CT, or retinal imaging, by learning measurement-specific representations from domain-specific quantitative measurements and leveraging these representations for downstream tasks.

Limitations. EchoTab requires specifying the mapping between diagnoses and the measurements relevant to them when constructing diagnosis-specific embeddings. While this mirrors clinical reasoning, it introduces domain knowledge into the model and requires specifying the relevant measurements when extending the framework to new diagnoses. In addition, we explore only a single strategy for combining the measurement-specific tokens relevant to each diagnosis, namely averaging them into a diagnosis-specific study embedding. More sophisticated aggregation mechanisms, such as learned weighting or attention, may better capture interactions among measurement tokens. Future work could therefore explore learning the measurement-diagnosis relationships, developing better aggregation strategies, and extending the framework to additional modalities and tasks.

AI USE STATEMENT

In this work, we used generative AI tools to design and provide feedback on our research methodology and experiments, implement methods, and interpret results. We have not used generative AI tools to generate synthetic datasets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, or clean and reformat datasets, and the remaining required disclosure tasks (formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis) are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, draft parts of a research paper, edit the paper to improve readability, and identify relevant literature. We have reviewed all AI-assisted work. In particular, AI feedback on our research methodology and experimental design was critically assessed by the authors, and final design decisions were made by the authors. AI-assisted implementations and code were verified and tested for correctness. AI-assisted interpretations of results were used as feedback rather than the final interpretation. All literature identified with AI assistance was read and verified by the authors to confirm its existence and relevance. All AI-drafted or AI-edited text was revised by the authors to ensure it accurately reflects our work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide our code at `https://anonymous.4open.science/r/EchoTab`, which includes the implementation of our model architecture and the downstream task. Our experiments use the MIMIC-IV-Echo dataset v1.0 (Gow et al., 2026; Pollard et al., 2026), which is distributed through PhysioNet under credentialed access. Researchers can obtain access by completing PhysioNet's credentialing process. We describe our data processing pipeline in Appendix A and the list of measurements we use in Appendix E. We report our model's hyperparameters in Appendix F and further details on the baselines in Appendix H.

REFERENCES

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

Rujikorn Charakorn, Edoardo Cetin, Yujin Tang, and Robert Tjarko Lange. Text-to-LoRA: Instant transformer adaption. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7485–7514. PMLR, 13–19 Jul 2025. URL `https://proceedings.mlr.press/v267/charakorn25a.html`.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21400–21413. Curran Associates, Inc., 2022. doi: 10.52202/068431-1555. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/867c06823281e506e8059f5c13a57f75-Paper-Conference.pdf`.

Brian Gow, Tom Pollard, Nathaniel Greenbaum, Benjamin Moody, Ahram Han, Jonathan W Waks, Alistair Johnson, Elizabeth Herbst, Parastou Eslami, Ashish Chaudhari, Tanner Carbonati, Seth Berkowitz, Roger Mark, and Steven Horng. MIMIC-IV-ECHO: Echocardiogram Matched Subset. *PhysioNet*, March 2026. doi: 10.13026/nrjh-5r77. URL `https://doi.org/10.13026/nrjh-5r77`. Version 1.0.

David Ha, Andrew M. Dai, and Quoc V. Le. Hypernetworks. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=rkpACe1lx`.

Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23924–23935, June 2023.

Marta Hasny, Maxime Di Folco, Keno Bressem, and Julia Schnabel. Tables guide vision: Learning to see the heart through tabular data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1716–1725, 2026.

Gregory Holste, Evangelos K. Oikonomou, Márton Tokodi, Attila Kovács, Zhangyang Wang, and Rohan Khera. Complete ai-enabled echocardiography interpretation with multitask deep learning. *JAMA*, 334(4):306–318, 07 2025. ISSN 0098-7484. doi: 10.1001/jama.2025.8731. URL https://doi.org/10.1001/jama.2025.8731.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d89a66c7c80a29b1bdbab0f2a1a94af8-Paper.pdf.

Myung Jun Kim, Félix Lefebvre, Gaëtan Brison, Alexandre Perez-Lebel, and Gaël Varoquaux. Table foundation models: on knowledge pre-training for tabular learning. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=QV4P8Csw17.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23q.html.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4804–4814, June 2022.

Alif Munim, Adibvafa Fallahpour, Teodora Szasz, Ahmadreza Attarpour, River Jiang, Brana Sooriyakanthan, Maala Sooriyakanthan, Heather Whitney, Jeremy Slivnick, Barry Rubin, Wendy Tsang, and Bo Wang. Echojepa: A latent predictive foundation model for echocardiography, 2026. URL https://arxiv.org/abs/2602.02603.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Milos Vukadinovic, I-Min Chiu, Xiu Tang, Neal Yuan, Tien-Yu Chen, Paul Cheng, Debiao Li, Susan Cheng, Bryan He, and David Ouyang. Comprehensive echocardiogram evaluation with view primed vision language ai. *Nature*, 650(8103):970–977, Feb 2026. ISSN 1476-4687. doi: 10.1038/s41586-025-09850-x. URL https://doi.org/10.1038/s41586-025-09850-x.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Ziyang Zhang, Qinxin Wu, Sirui Ding, Xiaolong Wang, and Jiancheng Ye. Echo-vision-fm: a pretraining and fine-tuning framework for echocardiogram video vision foundation model. *Nature Communications*, 17(1):19, Dec 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-66340-4. URL https://doi.org/10.1038/s41467-025-66340-4.

APPENDIX

A DATA PROCESSING

We construct paired echocardiogram-measurements datasets from two sources: an internal cohort and MIMIC-IV-Echo v1.0 (Gow et al., 2026; Pollard et al., 2026). For both, measurement selection and study filtering were performed independently, and patient-level splits were used throughout so that no patient appears in more than one split.

Measurement selection. Starting from all measurements available, we applied the filtering steps summarized in Table 4.

Table 4: Number of measurements remaining after each filtering step. Dashes indicate steps that do not apply to the given dataset.

Step	Internal	MIMIC
Initial measurements	647	302
Remove all-missing measurements	221	300
Remove stress-echo measurements	–	233
Remove non-measurement fields (demographics, hemodynamics)	–	195
Remove diagnostic labels (to be used for downstream tasks)	–	99
Remove all-missing measurements (second pass)	–	71
Remove measurements present in $<$100 studies	121	47
Remove redundant / merge synonymous measurements	67	45
Final	**67**	**45**

For our internal cohort, redundancy removal collapsed measurements of the same quantity obtained by different acquisition or computation methods, retaining a single preferred variant per quantity (preference order: biplane mode $>$ 4-chamber mode $>$ 2-chamber mode $>$ M-mode; 2D $>$ M-mode). We also excluded body-size-indexed measurements (e.g. LV mass index, stroke volume index), as these require anthropometric data that is not consistently available. For MIMIC, synonymous fields reporting identical quantities were merged (e.g., `biplane_lvef` and `lvef`). Measurements requiring M-mode acquisition were retained in both datasets.

Study filtering. Linking structured measurements to echo studies, excluding studies with no recorded values in the retained measurement list, and excluding studies without videos from the view set $\mathcal{C}$ (A2C, A3C, A4C, A5C, PLAX, PSAX, and Apical Doppler) yields **9,430 studies** for our internal cohort and **7,063 studies** for MIMIC.

Measurement pre-processing. Within each dataset, measurement values below the 1$^{\text{st}}$ or above the 99$^{\text{th}}$ percentile of the corresponding training-set distribution are set to missing. Measurement units are harmonized across datasets (e.g., milliseconds to seconds for deceleration and pressure half-times, and millimeters to centimeters for TAPSE).

Video sampling during training. Each study contributes at most one video per view type from the view set $\mathcal{C}$ (A2C, A3C, A4C, A5C, PLAX, PSAX, and Apical Doppler). During training, one video is randomly sampled from the available videos for each view type at every epoch, yielding a different combination of videos across epochs. At evaluation time, sampling is replaced by the first available video of each view type to ensure deterministic evaluation. For k-NN classification, the first available video of each view type is used regardless of training or evaluation mode.

Splits. Both datasets are split at the patient level into training, validation, and test sets (Table 5).

Table 5: Dataset splits, reported as number of patients and number of echo studies.

	Internal (patients / studies)	MIMIC (patients / studies)
Train	6,272 / 7,690	3,612 / 5,610
Validation	565 / 879	450 / 723
Test	566 / 861	452 / 730
Total	**7,403 / 9,430**	**4,514 / 7,063**

B Additional Results on Diagnosis Classification

B.1 Can a diagnosis-specific prior for the baselines close the gap to EchoTab?

EchoTab conditions its diagnosis predictions on diagnosis-relevant measurement tokens, and this conditioning is central to its performance: using all measurement tokens instead of diagnosis-relevant ones reduces its average AUC from 0.748 to 0.693 on MIMIC and from 0.778 to 0.751 on the internal dataset (Table 3). This raises the question of whether the advantage of EchoTab over the baselines stems solely from access to a diagnosis-specific prior. Since each measurement is acquired from specific echo views, the closest analogous prior available to the baselines is to restrict their input to the views relevant to each diagnosis. We therefore re-evaluate all baselines and echo foundation models using only diagnosis-relevant views (see Table 10 and Table 11). To ensure that both sides receive exactly one diagnosis-specific prior, EchoTab still uses all views, together with its diagnosis-relevant measurement tokens.

Table 6 shows that incorporating diagnosis-specific priors into the baselines via diagnosis-relevant views does not close the gap: EchoTab still achieves the highest average AUC on both datasets, outperforming the strongest baseline by 3.1 average AUC points on MIMIC (0.748 vs. 0.717 for MLP) and by 3.0 on the internal dataset (0.778 vs. 0.748 for study-level SupCon). This advantage is consistent across diagnoses: EchoTab achieves the best average AUC on 6 of 9 diagnoses on MIMIC, and on 7 of 8 diagnoses on the internal dataset. Together, this shows that the baselines cannot replicate the advantage of EchoTab, even when given an analogous prior through diagnosis-relevant views.

B.2 Does EchoTab benefit from diagnosis-relevant views as prior?

We next ask whether EchoTab itself benefits from diagnosis-specific views as prior. To this end, we restrict EchoTab to the same diagnosis-relevant views used for the baselines in Table 6, while keeping its diagnosis-relevant measurement tokens. As shown in Table 7, this lowers the average AUC from 0.748 to 0.725 on MIMIC and from 0.778 to 0.753 on the internal dataset. This suggests that the view-measurement relationships learned by the tabular decoder are more effective than a defined set of relevant views, either because the decoder weights the relevant views more effectively or because it extracts complementary information from views not considered relevant to a given diagnosis. In practice, EchoTab can therefore be applied directly to all views of a study, without any per-diagnosis view curation.

Views	Measurement tokens	MIMIC	Internal
All	Diagnosis-relevant	**.748**	**.778**
Diagnosis-relevant	Diagnosis-relevant	.725	.753

Table 7: **Ablation of the echo views used by EchoTab for diagnosis classification** (average AUC). Restricting EchoTab to diagnosis-relevant views lowers performance on both datasets, indicating that the tabular decoder benefits from access to all views rather than a defined set of relevant views.

C Dataset Details of Diagnosis Classification Tasks

We evaluate on a set of ordinal severity-grading tasks derived from the diagnostic labels in the structured measurements, which were held out during measurement selection. The labels in our internal cohort use a four-level scale (normal, mild, moderate, severe), while the labels in the MIMIC dataset use a seven-level scale that additionally distinguishes trace and the intermediate mild-moderate and

Table 6: **ECHOTAB still outperforms baselines when both are given diagnosis-specific prior.** Baselines and reference foundation models use only diagnosis-relevant views, whereas ECHOTAB uses all views with diagnosis-relevant measurement tokens. Bold and underlined values denote the **best** and second-best results.

(a) Diagnosis classification results on MIMIC (AUC).

Method	Echo encoder	Echo-tabular pre-training		Inference config.		MIMIC									
		Role of tabular data	Echo sample unit	Views	Measurement tokens	Valves			Heart dimension				Sys. function		Avg.
						AS	AR	MR	LA	RA	LV	RV	LVEF	RVEF	
Echo foundation models (Reference)															
PanEcho	PanEcho	–	–	Relevant	–	.561	**.643**	.615	.685	.693	.818	.673	.696	.720	.678
EchoJEPA-L	EchoJEPA-L	–	–	Relevant	–	.499	.490	.506	.615	.710	.656	.679	.609	.697	.607
EchoPrime	EchoPrime	–	–	Relevant	–	.552	.511	.583	.695	.718	.749	.667	.710	.742	.659
Baselines															
Contrastive	EchoPrime	Tabular alignment target	view	Relevant	–	.603	.578	.627	**.748**	.755	.783	.684	.709	.752	.693
Contrastive	EchoPrime	Tabular alignment target	study	Relevant	–	.671	.568	.612	.717	.762	.796	.722	.728	.768	.705
SupCon	EchoPrime	Tabular guidance	view	Relevant	–	.664	.548	.630	.731	.743	.720	.690	.812	**.771**	.701
SupCon	EchoPrime	Tabular guidance	study	Relevant	–	.635	.514	.652	.719	.709	.746	.714	.750	.745	.687
MLP	EchoPrime	Tabular reconstruction	study	Relevant	–	.664	.591	.614	.734	.721	.797	.719	**.843**	**.771**	.717
Ours															
ECHOTAB	EchoPrime	Tabular reconstruction	study	All	Relevant	**.723**	.592	**.658**	.745	**.808**	**.852**	**.756**	**.843**	.754	**.748**

(b) Diagnosis classification results on the internal dataset (AUC).

Method	Echo encoder	Echo-tabular pre-training		Inference config.		Internal								
		Role of tabular data	Echo sample unit	Views	Measurement tokens	Valves		Heart dimension				Sys. function		Avg.
						AS	MR	LA	RA	LV	RV	LVEF	RVEF	
Echo foundation models (Reference)														
PanEcho	PanEcho	–	–	Relevant	–	.681	.694	.706	.724	.786	.664	.724	.618	.700
EchoJEPA-L	EchoJEPA-L	–	–	Relevant	–	.612	.593	.674	.662	.563	.615	.670	.744	.642
EchoPrime	EchoPrime	–	–	Relevant	–	.753	.681	.704	.767	.770	.626	.748	.656	.713
Baselines														
Contrastive	EchoPrime	Tabular alignment target	view	Relevant	–	.762	.667	.752	.759	.774	.679	.766	.661	.728
Contrastive	EchoPrime	Tabular alignment target	study	Relevant	–	.790	.692	.750	.739	.787	.651	.757	.644	.726
SupCon	EchoPrime	Tabular guidance	view	Relevant	–	.793	.629	.763	**.769**	.764	.719	.790	.685	.739
SupCon	EchoPrime	Tabular guidance	study	Relevant	–	.797	.705	.784	.746	.829	.734	.786	.606	.748
MLP	EchoPrime	Tabular reconstruction	study	Relevant	–	.776	.708	.783	.727	.799	.676	.813	.689	.746
Ours														
ECHOTAB	EchoPrime	Tabular reconstruction	study	All	Relevant	**.800**	**.713**	**.787**	.766	**.830**	**.749**	**.829**	**.749**	**.778**

moderate-severe grades. Studies without a label for a given task are excluded from that task, so per-task totals differ within a split.

Table 8 and Table 9 give the class distributions for MIMIC and our internal cohort, respectively. For training, we balance the classes as much as possible so that the k-NN classifier sees roughly an equal number of samples from each class when making predictions. The validation and test sets retain the original class distributions. Meanwhile, Table 10 and Table 11 show the diagnosis-to-measurement-to-view mappings for MIMIC and our internal dataset, respectively, as defined by a cardiologist and used to identify diagnosis-relevant measurements and views.

D ECHOS

D.1 ECHO STUDY DISTRIBUTION

Figure 4 shows the distribution of echo studies by echo view and data split (train/val/test) for the MIMIC dataset and our internal cohort.

D.2 ECHO PRE-PROCESSING

Each echo video first undergoes pre-processing to isolate the ultrasound region using the `SequenceOfUltrasoundRegions` field from the DICOM metadata, and to strip out the ECG overlay along with any remaining burned-in text or annotations. The videos are then resized to 224×224 pixels and stored as AVI files. As in EchoPrime, an additional 10% zoom crop is ap-

Table 8: Class distribution of the downstream diagnosis classification tasks of the MIMIC dataset (*N*ormal, *T*race, *Mi*ld, *Mi-Mo* mild-to-moderate, *Mo*derate, *Mo-S* moderate-to-severe, *S*evere).

Task	N	T	Mi	Mi-Mo	Mo	Mo-S	S
Train							
Aortic stenosis (AS)	115	-	112	-	114	-	112
Aortic regurgitation (AR)	-	9	8	9	9	9	9
Mitral regurgitation (MR)	73	75	74	75	74	73	74
Left atrium size (LA)	417	-	413	-	417	-	-
Right atrium size (RA)	152	-	153	-	152	-	149
Left ventricle size (LV)	5	-	5	-	5	-	5
Right ventricle size (RV)	29	-	29	-	29	-	29
Left ventricle ejection fraction (LVEF)	40	-	38	-	38	-	40
Right ventricle ejection fraction (RVEF)	87	-	87	-	85	-	86
Val							
Aortic stenosis (AS)	12	-	22	-	20	-	22
Aortic regurgitation (AR)	-	136	102	20	12	2	-
Mitral regurgitation (MR)	9	300	195	61	56	25	7
Left atrium size (LA)	243	-	72	-	58	-	-
Right atrium size (RA)	325	-	177	-	38	-	17
Left ventricle size (LV)	175	-	9	-	8	-	-
Right ventricle size (RV)	531	-	100	-	17	-	4
Left ventricle ejection fraction (LVEF)	100	-	31	-	2	-	10
Right ventricle ejection fraction (RVEF)	577	-	53	-	24	-	15
Test							
Aortic stenosis (AS)	20	-	33	-	9	-	19
Aortic regurgitation (AR)	-	134	114	24	14	1	2
Mitral regurgitation (MR)	7	338	176	61	54	30	5
Left atrium size (LA)	254	-	92	-	50	-	-
Right atrium size (RA)	346	-	172	-	33	-	23
Left ventricle size (LV)	216	-	5	-	14	-	3
Right ventricle size (RV)	508	-	113	-	41	-	2
Left ventricle ejection fraction (LVEF)	78	-	40	-	5	-	3
Right ventricle ejection fraction (RVEF)	555	-	67	-	35	-	16

Table 9: Class distribution of the downstream diagnosis classification tasks of our internal dataset (*N*ormal, *Mi*ld, *Mo*derate, *S*evere).

	Train				Val				Test			
Task	N	Mi	Mo	S	N	Mi	Mo	S	N	Mi	Mo	S
Aortic stenosis (AS)	96	96	96	96	724	28	21	10	736	21	22	20
Mitral regurgitation (MR)	37	37	37	37	557	176	40	10	552	202	35	10
Left atrium size (LA)	113	113	113	113	626	47	79	26	658	46	58	24
Right atrium size (RA)	73	73	73	73	704	29	27	19	738	15	21	17
Left ventricle size (LV)	22	22	22	22	714	12	19	9	711	14	14	6
Right ventricle size (RV)	14	14	14	14	721	26	8	-	716	17	12	-
Left ventricle ejection fraction (LVEF)	109	109	109	109	633	62	53	43	634	71	55	47
Right ventricle ejection fraction (RVEF)	13	13	13	13	751	22	6	3	770	16	11	-

plied at data-loading time to cut down on empty background, after which the frames are resized back to 224×224. Pixel values are normalized based on the mean and standard deviation from the EchoPrime training set. Each video is also sampled down to 16 frames using a stride of 2. Videos that come up short are zero-padded to a uniform length. During training, no augmentation is applied to the echos.

D.3 ECHO TOKEN SEQUENCE

Table 12 reports k-NN ($k = 5$) performance on the view classification task across different configurations of the echo tokens from EchoPrime. Extracting local patch tokens from EchoPrime and subsequently removing those corresponding to background patches yields the best performance, indicating that this configuration produces the most discriminative echo tokens across views. This is a desirable property because we want the measurement queries in ECHOTAB to learn to attend to the correct echo views.

Table 10: Diagnosis → Measurement → View mapping for the MIMIC dataset, as defined by a cardiologist.

Diagnosis	Measurements	Views
Aortic regurgitation (AR)	ar_pr_halftime	Apical Doppler
Aortic stenosis (AS)	av_mean_grad, av_pk_grad, av_pk_vel, av_area_continuity, av_vti, lvot_diam, lvot_pk_vel, lvot_vti	Apical Doppler, PLAX
Mitral regurgitation (MR)	mv_e_decel	Apical Doppler
Left atrium size (LA)	la_dimen, la_4ch_length, la_vol	PLAX, A4C, A2C
Right atrium size (RA)	ra_length	A4C
Left ventricle size (LV)	lvedd, lvesd, biplane_lvedv, biplane_lvesv	PLAX, A2C, A4C
Right ventricle size (RV)	rv_diam	A4C
Left ventricle ejection fraction (LVEF)	biplane_lvef	A2C, A4C
Right ventricle ejection fraction (RVEF)	tapse	A4C

Table 11: Diagnosis → Measurement → View mapping for our internal dataset, as defined by a cardiologist.

Diagnosis	Measurements	Views
Aortic stenosis (AS)	AV_Pmean, AV_Pmax, AV_Vmax, AV_VTI, AV_COA_Vmax, LVOT_Pmax, LVOT_Pmean, LVOT_VTI, LVOT_SV_D, LVOT_Vmax, LVOT_Vmean, LVOTd_2D	Apical Doppler, PLAX
Mitral regurgitation (MR)	MV_DT	Apical Doppler
Left atrium size (LA)	LADs_AC4, LAESV_MODBP, LAAs_MOD4	PLAX, A4C, A2C
Right atrium size (RA)	RAA_2D, RAD_l_2D, RAV_A4C	A4C
Left ventricle size (LV)	LVDd_2D, LVDs_2D, LVEDV_MODBP, LVESV_MODBP	PLAX, A2C, A4C
Right ventricle size (RV)	RVDd_2D	A4C
Left ventricle ejection fraction (LVEF)	LVEF_MODBP, LVSV_MODBP	A2C, A4C
Right ventricle ejection fraction (RVEF)	TAPSE	A4C

E MEASUREMENTS

Figure 5 and Figure 6 show the list of measurements used from the MIMIC dataset and our internal dataset, respectively.

F HYPERPARAMETER SETTINGS

ECHOTAB is trained for 30 epochs using the AdamW optimizer with a base batch size of 4 and gradient accumulation over 16 iterations. d is set to 768. We use a base learning rate of 1×10^{-4} for the EchoPrime-initialized echo encoder on MIMIC and 5×10^{-5} on our internal dataset, while using 1×10^{-4} for all other trainable parameters, including the tabular decoder, measurement queries, and prediction head. We use a warmup-cosine schedule, linearly increasing the learning rate from 10% to 100% of the base rate over two epochs, followed by cosine decay to 10% by the end of training. The tabular decoder consists of two cross-attention blocks with four attention heads each, while all prediction-head variants use a two-layer MLP. For the partially shared head, we use a low-rank update with rank $r = 8$ and scaling factor $\alpha = 16$. We apply gradient clipping with a maximum norm of 50.

G HYPERNETWORK VARIANTS

Inspired by text-to-LoRA (Charakorn et al., 2025), we experimented with three variants of the hypernetwork used in the partially shared prediction head of ECHOTAB: `L`, `M`, and `S`. By default, our model uses `M`.

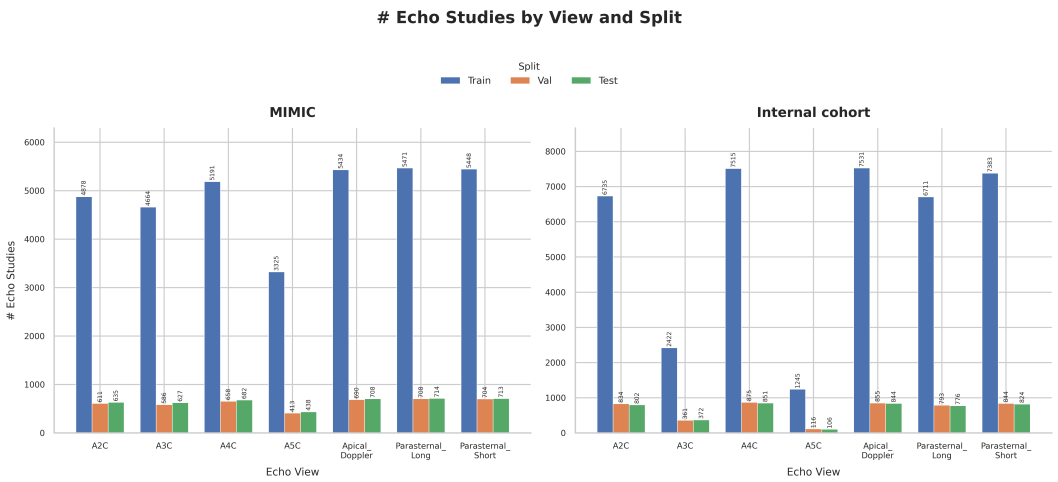


Figure 4: Distribution of echo studies by view and data split (train/val/test) for the MIMIC and internal cohort datasets.

Token type	Remove background tokens	AUROC	Accuracy
Global (CLS)	-	0.752	0.456
Local	×	0.792	0.554
Local	✓	**0.805**	**0.583**

Table 12: k**-NN** ($k = 5$) **performance on echo view classification using different configurations of echo tokens from EchoPrime.** We use the configuration in the last row for all models that use EchoPrime: local tokens, excluding those that correspond to the background.

All variants share the same hypernetwork backbone (Figure 7) and differ only in how they generate the LoRA factors $A_{m,\ell} \in \mathbb{R}^{d\times r}$ and $B_{m,\ell} \in \mathbb{R}^{r\times d}$, with $d = 768$ and $r = 8$ (Figure 8). The largest variant, `L`, generates both factors A and B in a single forward pass, requiring an output dim. of $2rd = 12288$ in `h1w`. The medium variant, `M`, instead runs the shared body twice, generating one factor per pass and reducing the output dim. of `h1w` to $rd = 6144$. The two passes are distinguished by a learned `ab_emb` embedding, with one embedding vector for each factor, which is added to the conditioning vector c. The smallest variant, `S`, further factorizes generation by producing one rank slice of a single factor at a time. It therefore reduces the output dimension to $d = 768$ while requiring $2r = 16$ forward passes through the shared body. Each pass is conditioned on both the corresponding `ab_emb` and a learned `rank_emb`, which provides one embedding vector for each rank index. Thus, moving from `L` to `S` trades additional computation for fewer hypernetwork parameters: the number of parameters decreases, while the number of forward passes through the shared body increases. Table 13 shows the ablation of these different hypernetwork variants.

Hypernetwork variant	Diag. clf. (avg. AUC)	Meas. pred. (avg. R^2)
`L`	.735	.405
`M`	**.748**	**.418**
`S`	.743	.395

Table 13: **Ablation of the hypernetwork variants used in the partially shared prediction head of ECHOTAB on the MIMIC dataset.** By default, our model uses the `M` variant.

```
ar_pr_halftime, arch_diam, ascending_diam, av_area_continuity,
av_mean_grad, av_pk_grad, av_pk_vel, av_vti, biplane_lvedv,
biplane_lvef, biplane_lvesv, descending_diam, e_e_prime, fs,
inf_lat_thickness, ivc_diam, la_4ch_length, la_dimen, la_vol, lat_e_prime,
lateral_E_to_eprime, lvedd, lvesd, lvot_diam, lvot_pk_vel, lvot_vti,
mean_E_to_eprime, mv_e_decel, mv_mean_grad, mv_peak_a, mv_peak_e,
mv_peak_e_a, mv_peak_vel, mva2, pv_peak_vel, ra_length, rest_lvot, rv_diam,
sept_e_prime, septal_E_to_eprime, septal_thickness, sinus_diam, tapse,
tr_mmhg, tr_velocity
```

Figure 5: List of measurements used from the MIMIC dataset (45 total).

```
AoAsc_2D, Ao_MM, AR_DS, AR_PHT, AR_Pmax, AR_Vmax, AVA_D, AV_COA_Vmax,
AV_Pmax, AV_Pmean, AV_VR, AV_VTI, AV_Vmax, AV_Vmean, IVSd_2D, LAAs_MOD4,
LADs_AC4, LAESV_MODBP, LA_AO_MM, FS_2D, IVS_LVPW_T, LVAd_AL4, LVAs_AC4_2D,
LVDd_2D, LVDs_2D, LVEDV_MODBP, LVEF_MODBP, LVESV_MODBP, LVLd_AL4, LVLs_AL4,
LVM_MM, LVOTA_2D, LVOT_Pmax, LVOT_Pmean, LVOT_SV_D, LVOT_VTI, LVOT_Vmax,
LVOT_Vmean, LVOTd_2D, LVPWd_2D, LVSV_MODBP, MVA_PHT, MV_A_Vmax, MV_DS,
MV_DT, MV_EA, MV_E_Eann, MV_E_Eann_lat, MV_E_Vmax, MV_Eann, MV_Eann_lat,
MV_PHT, MV_Pmax, MV_Pmean, MV_VTI, MV_Vmax, MV_Vmean, PV_ACCS, PV_Acctime,
RAA_2D, RAD_l_2D, RAV_A4C, RVDd_2D, RV_LV_EDD, TAPSE, TR_Pmax, TR_Vmax
```

Figure 6: List of measurements used from our internal dataset (67 total).

H BASELINES

This section describes the baselines used in our experiments in more detail: MLP (Section H.1), contrastive (Section H.2), and SupCon (Section H.3). All baselines use the same EchoPrime-initialized echo encoder and view set $\mathcal{C}$ as ECHOTAB.

H.1 MLP

The MLP baseline adds a separate prediction head for each measurement on top of the echo encoder, analogous to the independent-head variant of ECHOTAB. The study-level echo embedding is constructed by averaging over the echo view embeddings. It is then passed to each measurement-specific head, with all heads using the same L-layer MLP architecture as ECHOTAB's prediction head.

The MLP baseline is trained for 30 epochs with the AdamW optimizer, using a base batch size of 4 and gradient accumulation over 16 iterations. We use a base learning rate of 1×10^{-4} for both the MIMIC and internal datasets. We use a warmup-cosine schedule, linearly increasing the learning rate from 10% to 100% of the base rate over two epochs, followed by cosine decay to 10% by the end of training. For the echo encoder, we apply a layer-wise learning-rate decay of 0.75 on MIMIC and no decay on the internal dataset.

H.2 CONTRASTIVE

We align paired echo and measurement representations using a CLIP-style contrastive objective (Radford et al., 2021), following the image–tabular formulation of MMCL (Hager et al., 2023). For a batch of N samples, let $\mathbf{e}_i$ denote the echo representation and $\mathbf{t}_i$ the corresponding measurement representation. Both are projected to 128-dim by separate linear projection layers, yielding $\mathbf{z}_i^{\mathrm{e}}$ and $\mathbf{z}_i^{\mathrm{t}}$, respectively. The echo-to-measurement contrastive loss is

$$\mathcal{L}_{\mathrm{E}\to\mathrm{M}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(\cos(\mathbf{z}_i^{\mathrm{e}}, \mathbf{z}_i^{\mathrm{t}})/\tau\right)}{\sum_{j=1}^{N} \exp\left(\cos(\mathbf{z}_i^{\mathrm{e}}, \mathbf{z}_j^{\mathrm{t}})/\tau\right)}, \tag{7}$$

```
Hypernet(
  (cond_projection): Sequential(
    (0): Linear(in_features=768, out_features=128, bias=True)
    (1): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
  )
  (layer_idx_emb): Embedding(1, 128)
  (layer_idx_emb_norm): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
  (body): Sequential(
    (0): Linear(in_features=128, out_features=256, bias=True)
    (1): GELU()
    (2): Dropout(p=0.05, inplace=False)
    (3): Linear(in_features=256, out_features=128, bias=True)
    (4): GELU()
  )
  (h1w_head): Linear(in_features=128, out_features=6144, bias=True)
  (h1b_head): Linear(in_features=128, out_features=768, bias=True)
  (h2w_head): Linear(in_features=128, out_features=768, bias=True)
  (h2b_head): Linear(in_features=128, out_features=1, bias=True)
)
```

Figure 7: Hypernetwork that generates the measurement-specific corrections in the partially shared prediction head (pictured here is the `M` variant). The d-d measurement query is projected to a 128-d conditioning vector c, passed through the body ($h = \mathrm{body}(c) + c$), and decoded by four linear heads: `h1w`/`h1b` produce the rank-8 LoRA factors ($A_{m,\ell}$ and $B_{m,\ell}$) and bias ($\beta_{m,\ell}$) of the hidden layer ℓ of shape $d \times d$, and `h2w`/`h2b` the weight delta ($a_{m,L}$) and bias ($\beta'_{m,L}$) of the final layer L of shape $d \times 1$. Additionally, the `h1` heads are conditioned on $c + \mathrm{LN}(\texttt{layer_idx_emb}[\ell])$, so that it can generate a distinct hidden layer for each depth ℓ. The `h2` heads use c directly, as there is only a single final layer L. All generated parameters are added to the base MLP shared across measurements.

```
# L
(h1w_head): Linear(in_features=128, out_features=12288, bias=True)

# M
(ab_emb):      Parameter (2, 128)
(ab_emb_norm): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
(h1w_head):    Linear(in_features=128, out_features=6144, bias=True)

# S
(ab_emb):        Parameter (2, 128)
(ab_emb_norm):   LayerNorm((128,), eps=1e-05, elementwise_affine=True)
(rank_emb):      Parameter (8, 128)
(rank_emb_norm): LayerNorm((128,), eps=1e-05, elementwise_affine=True)
(h1w_head):      Linear(in_features=128, out_features=768, bias=True)
```

Hypernetwork variant	# passes	`h1w_head` output dim.
`L`	1	$2rd = 12288$
`M`	2	$rd = 6144$
`S`	$2r = 16$	$d = 768$

Figure 8: Hypernetwork variants used in the partially shared prediction head of ECHOTAB. Assume $r = 8$ and $d = 768$. Every other module of the hypernetwork is identical across variants (Figure 7).

where τ is the temperature. We set $\tau = 0.07$. We additionally compute the symmetric measurement-to-echo loss $\mathcal{L}_{\mathrm{M}\to\mathrm{E}}$ by exchanging the two modalities, and optimize a weighted combination of the two directional losses.

We consider two variants that differ in the sample unit. In the *study-level* variant, $\mathbf{e}_i$ is obtained by attention-pooling the available view representations of a study. In the *view-level* variant, $\mathbf{e}_i$ is a view representation and is paired with the measurement representation of its corresponding study. The view-level variant therefore provides substantially more samples, enabling larger batches and more negative pairs.

The measurements for each study are encoded with the TARTE tabular encoder (Kim et al., 2025). Because TARTE requires column names corresponding to the input values, we use the measurement

descriptions provided by MIMIC and our internal dataset as column names. We use the TARTE `CLS` token as $\mathbf{t}_i$. To maximize batch size, these measurement representations are pre-computed using a pre-trained and frozen TARTE encoder, while the echo encoder is fully fine-tuned.

We train for 30 epochs using the AdamW optimizer with a batch size of 192 for the view-level variant and 32 for the study-level variant. The base learning rate is 5×10^{-5} for the study-level variant on both datasets and 1×10^{-4} for the view-level variant on both datasets. We use a warmup-cosine schedule, linearly increasing the learning rate from 10% to 100% of the base rate over two epochs, followed by cosine decay to 10% by the end of training.

H.3 Supervised Contrastive (SupCon)

The SupCon baseline uses tabular measurement similarity to define positive and negative echo pairs, following TGV (Hasny et al., 2026) and inspired by the original SupCon (Khosla et al., 2020). Each measurement is z-score normalized using training-set statistics, and missing values are mean-imputed, corresponding to zero after normalization.

For a batch of N samples, let $\mathbf{e}_i$ denote the echo representation after projection by a linear layer and $\mathbf{x}_i \in \mathbb{R}^F$ the corresponding vector of F normalized measurements. We ℓ_2-normalize the representations:

$$\hat{\mathbf{e}}_i = \frac{\mathbf{e}_i}{|\mathbf{e}_i|_2}, \qquad \hat{\mathbf{x}}_i = \frac{\mathbf{x}_i}{|\mathbf{x}_i|_2}. \tag{8}$$

The tabular similarity between samples i and j is then

$$S_{ij} = \hat{\mathbf{x}}_i^\top \hat{\mathbf{x}}_j. \tag{9}$$

Following TGV, positives are defined relative to each anchor rather than using a fixed similarity threshold. For anchor i, let

$$M_i = \max_{j \neq i} S_{ij}, \tag{10}$$

and define the positive set as

$$\mathcal{P}(i) = \{\, j \neq i \mid S_{ij} \geq M_i - h \,\}. \tag{11}$$

which guarantees at least one positive for every anchor. All remaining samples in the batch serve as negatives. The training objective is

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\sum_{j \in \mathcal{P}(i)} \exp\left(\hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_j / \tau\right)}{\sum_{k=1}^{N} \exp\left(\hat{\mathbf{e}}_i^\top \hat{\mathbf{e}}_k / \tau\right)}, \tag{12}$$

where τ is the temperature. We set $\tau = 0.07$ and $h = 0.05$.

As in the contrastive baseline, we consider both *study-level* and *view-level* variants, using the same definitions of the sample unit. We train for 30 epochs using the AdamW optimizer with a batch size of 192 for the view-level variant and 32 for the study-level variant. The base learning rate is 1×10^{-4} for the study-level variant and 1×10^{-4} for the view-level variant on MIMIC, while it is 5×10^{-5} for both variants on our internal dataset. We use a warmup-cosine schedule, linearly increasing the learning rate from 10% to 100% of the base rate over two epochs, followed by cosine decay to 10% by the end of training.

I Detailed Results on Measurement Prediction

Figure 9 and Figure 10 show the per-measurement and per-section R^2 by EchoTab with a partially shared head on the MIMIC dataset.

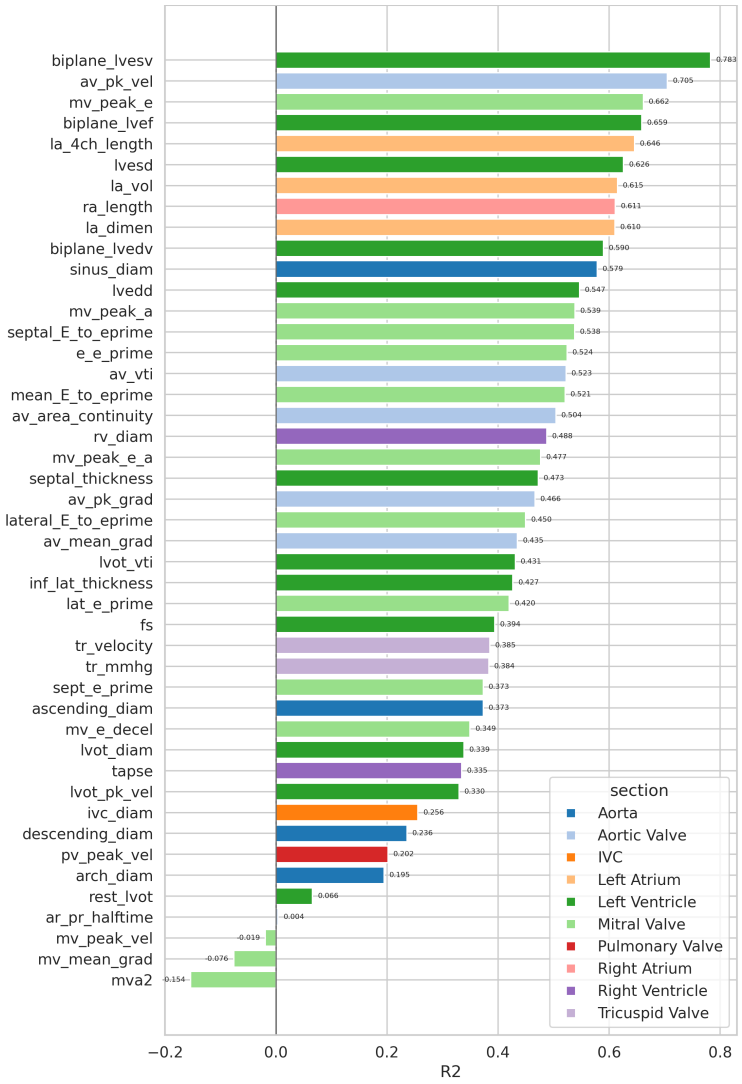


Figure 9: Per-measurement R^2 by ECHOTAB on the MIMIC dataset.

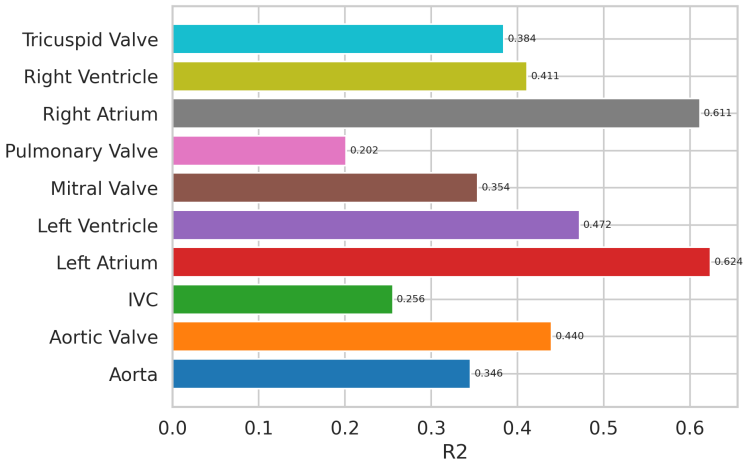


Figure 10: Per-section R^2 by ECHOTAB on the MIMIC dataset.